Test Big Data Samples

BSBXBD402

Lawrence Lim

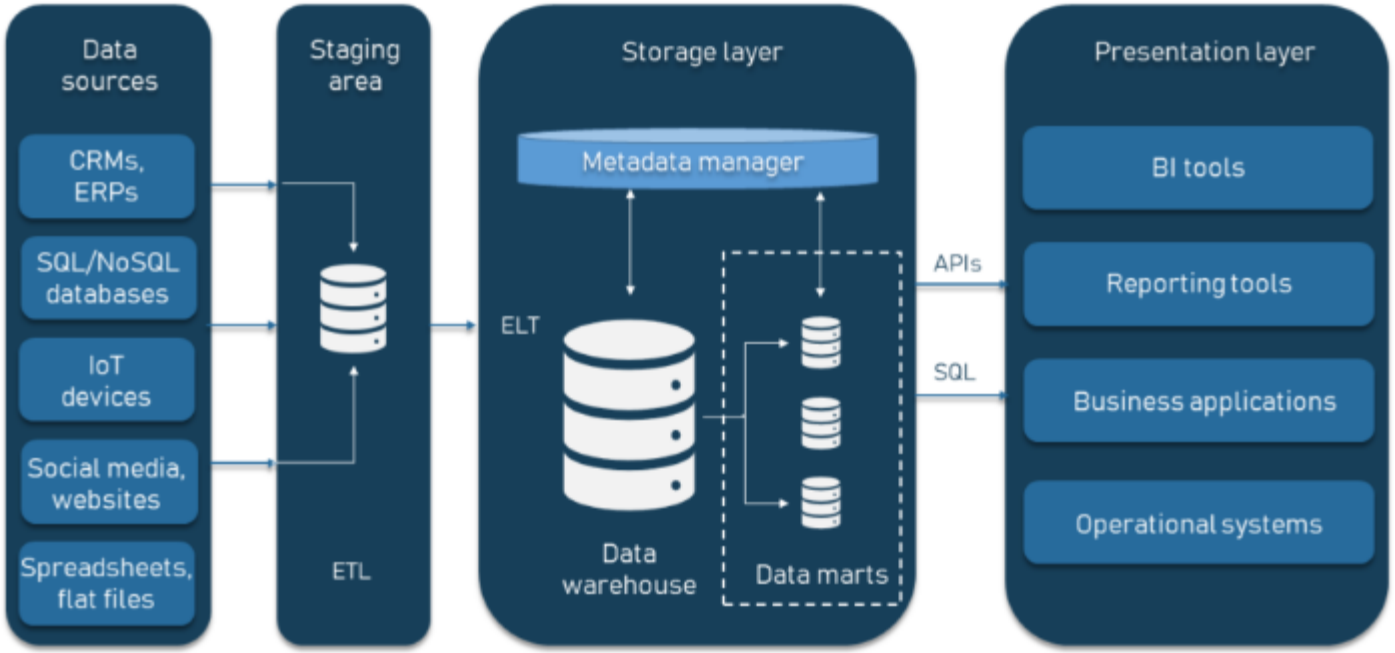
Lawrence.lim@cdu.edu.au

Term 3 2023

CHARLES DARWIN UNIVERSITY AUSTRALIA

Charles Darwin University acknowledges all First Nations people across the lands on which we live and work, and we pay our respects to Elders both past and present.

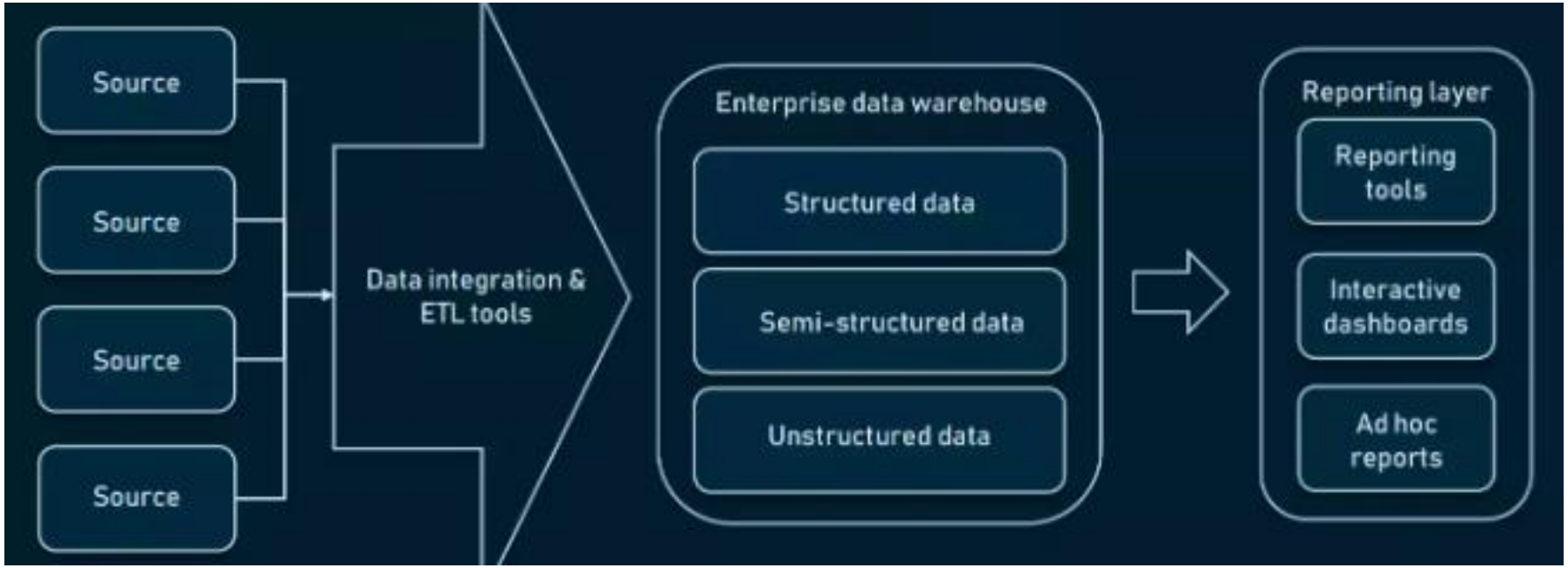
Enterprise Data Warehouse (EDW) components



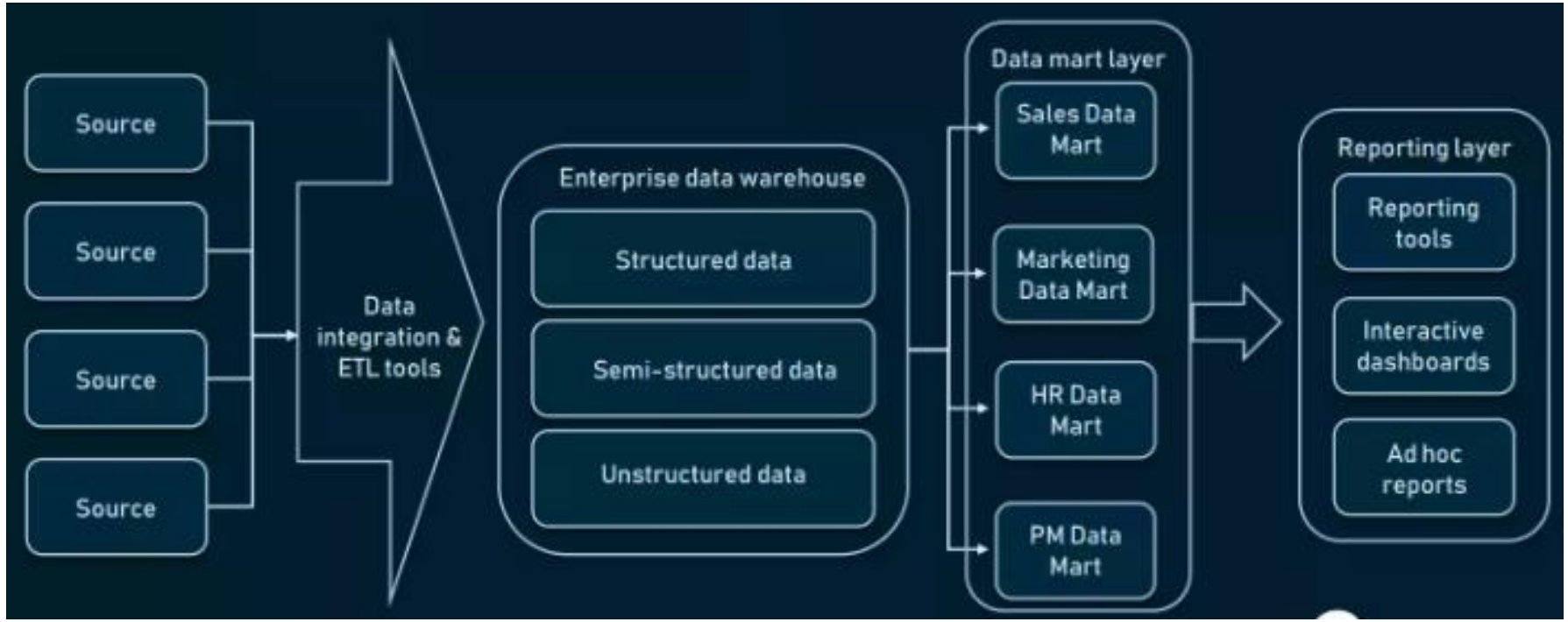
EDW attributes

- An EDW provides the likeness of the original data source in a single repository
- Data stored in an EDW is always standardised and structured
- Subject–oriented: e.g. sales data model
- Time-dependent: collected data usually historical data
- Non-volatile: data is never deleted from EDW

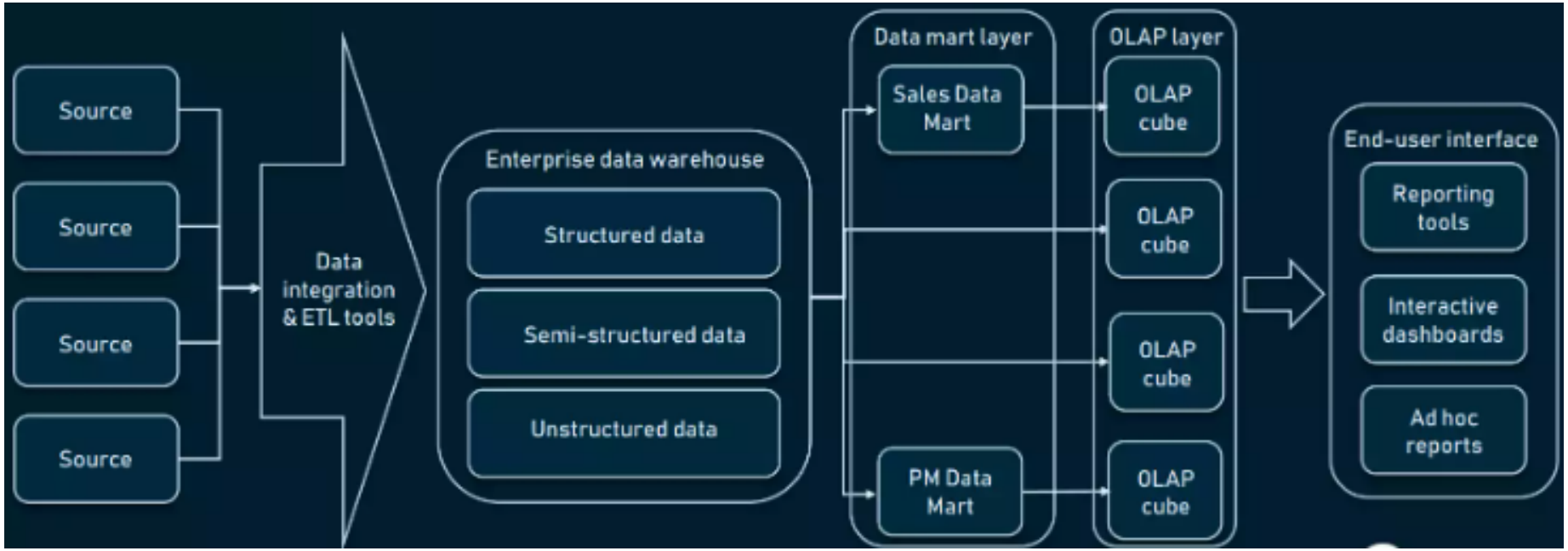
1-tier DW Architecture



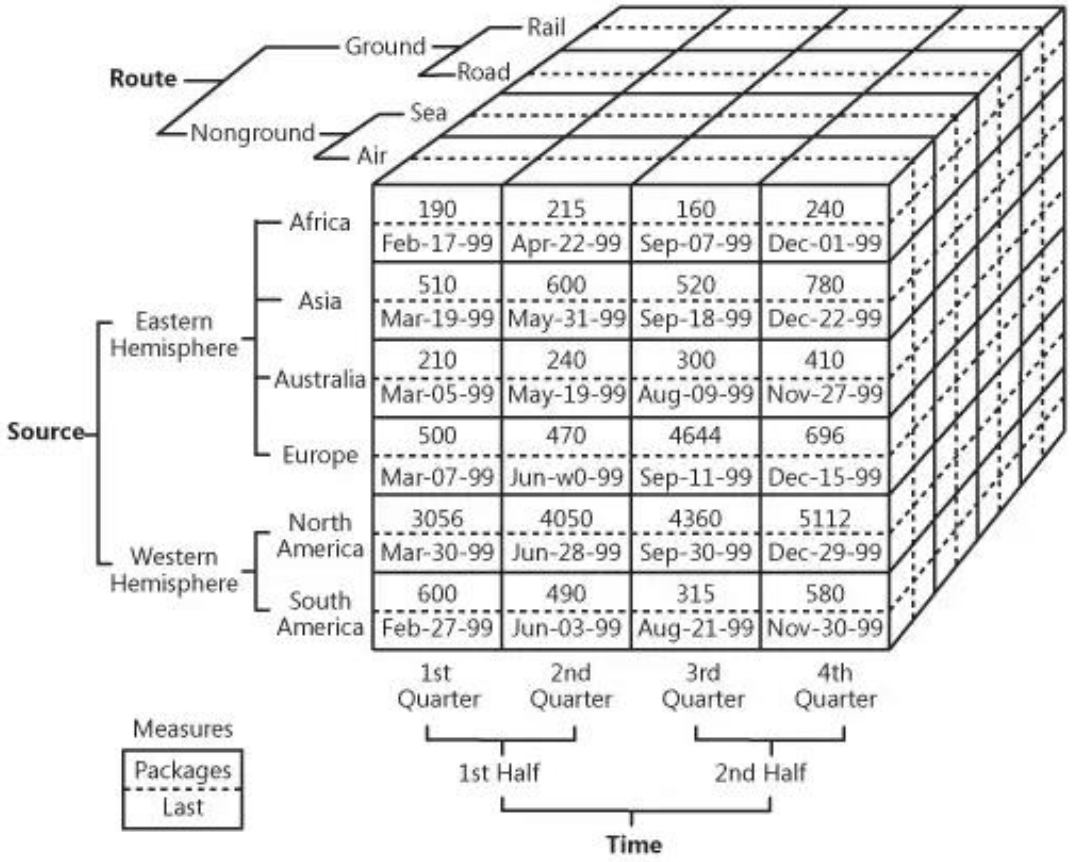
2-tier DW Architecture



3-tier DW Architecture



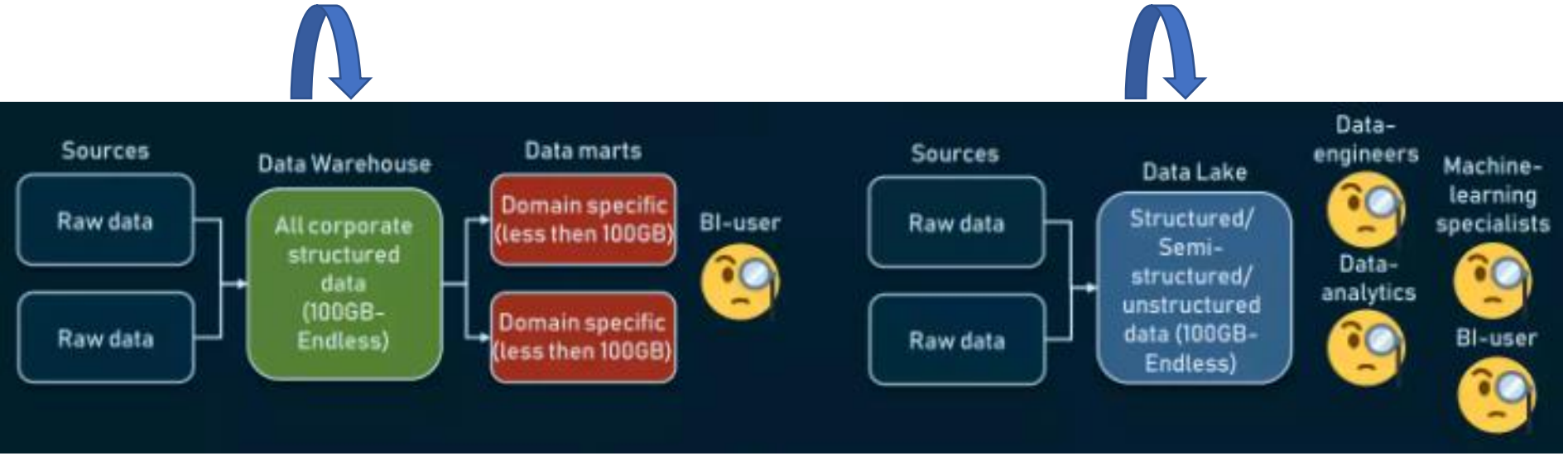
OnLine Analytical Processing (OLAP) cube



DATA WAREHOUSES VS DATA LAKES VS DATA MARTS

	Data warehouses	Data lakes	Data marts
Usage	The data analysis and reporting needs of an entire organization	The reporting needs of different kinds and difficulty, predictive analytics	The reporting needs of a specific operational department or subject
Data stored (typically)	Larger volumes of structured data; processed	Huge volumes of structured and unstructured data; raw	A limited amount of structured data; processed
Data sources	An array of external and internal sources, covering different areas of business	Any external or internal sources	Few sources linked to one business area
Size	Larger than 100 GB	Larger than 100 GB	Smaller than 100 GB
Ease of creation	Difficult to set up	Difficult to set up	Easy to set up

DW vs Data marts vs Data lakes



EDW technologies

- Recently, cloud technologies is becoming more popular
- Providers offer dw-as-a-service (DaaS) – provides computational power, storage, resource and server management, BI tools
- Pricing based on access per hour, number of concurrent users, data storage size, etc.
- Popular ones are:
  - Amazon Redshift (more self-managed – need own data engineers)
  - GoogleBigQuery (serverless technology – management taken care of)
  - Snowflake (serverless)

Different forms of Pre-analysis
Important tasks for Data analysts

- Data cleaning
- Data transformation
- Data integration
- Data reduction
- Data discretisation

Data Cleaning

Why cleaning Is required?

- **Incomplete.** When some of the attribute values are lacking, certain attributes of interest are lacking, or attributes contain only aggregate data.
- **Noisy.** When data contains errors or outliers. For example, some of the data points in a dataset may contain extreme values that can severely affect the dataset's range.
- **Inconsistent.** Data contains discrepancies in codes or names. For example, if the "Name" column for registration records of employees contains values other than alphabetical letters, or if records do not start with a capital letter, discrepancies are present.

Data Cleaning: 3 methods to “clean” “dirty” data

Clean → better organised, scrubbed of incorrect, incomplete, or duplicated data

1. **Data Munging (or Wrangling) – turning data to be easier understood**

Consider “Add two diced tomatoes, three cloves of garlic, and a pinch of salt in the mix”

After wrangling >>>

Ingredient	Quantity	Unit/size
Tomato	2	Diced
Garlic	3	Cloves
Salt	1	Pinch

Data Cleaning: 3 methods to “clean” “dirty” data

Cleaning → better organised, scrubbed of incorrect, incomplete, or duplicated data

2. Handling Missing Data

Sometimes data may be in the right format, but some of the values are missing. Consider a table containing customer data in which some of the home phone numbers are absent. This could be due to the fact that some people do not have home phones – instead they use their mobile phones as their primary or only phone.

Furthermore, some data may get lost due to system or human error while storing or transferring the data.

So, what to do when encountering missing data?

Data Cleaning: 3 methods to “clean” “dirty” data

Cleaning → better organised, scrubbed of incorrect, incomplete, or duplicated data

3. **Smooth Noisy Data**
   - **Times when data is not missing but corrupted**
   - **Data corruption caused by faulty data collection instruments and/or data entry problems**

Fix??

- **Identify / remove outliers**
- **Resolve inconsistencies in the data**

#	Country	Alcohol	Deaths	Heart	Liver
1	Australia	2.5	785	211	15.30000019
2	Austria	3.000000095	863	167	45.59999847
3	Belg. and Lux.	2.900000095	883	131	20.70000076
4	Canada	2.400000095	793	NA	16.39999962
5	Denmark	2.900000095	971	220	23.89999962
6	Finland	0.800000012	970	297	19
7	France	9.100000381	751	11	37.90000153
8	Iceland	−0.800000012	743	211	11.19999981
9	Ireland	0.699999988	1000	300	6.5
10	Israel	0.600000024	−834	183	13.69999981
11	Italy	27.900000095	775	107	42.20000076
12	Japan	1.5	680	36	23.20000076
13	Netherlands	1.799999952	773	167	9.199999809
14	New Zealand	1.899999976	916	266	7.699999809
15	Norway	0.0800000012	806	227	12.19999981
16	Spain	6.5	724	NA	NA
17	Sweden	1.600000024	743	207	11.19999981
18	Switzerland	5.800000191	693	115	20.29999924
19	UK	1.299999952	941	285	10.30000019
20	US	1.200000048	926	199	22.10000038
21	West Germany	2.700000048	861	172	36.70000076

Excessive wine Consumption and mortality rate

Excessive wine Consumption and mortality rate

a. Name of the country from which sample obtained
b. Alcohol consumption measured as liters of wine, per capita
c. Number of deaths from alcohol consumption, per 100,000 people
d. Number of heart disease deaths, per 100,000 people
e. Number of deaths from liver diseases, also per 100,000 people

Noisy data?

Missing data?

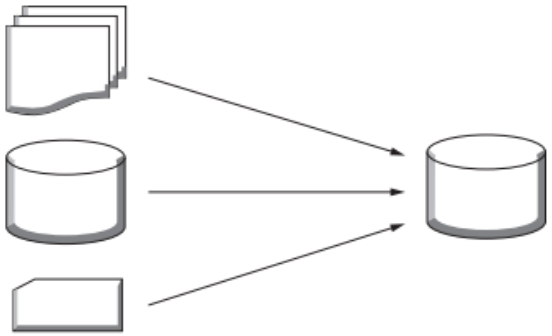
Data wrangling?

Data Integration

Another data source which is about alcohol consumption and number of related fatalities across seven States in India

Integrate India's attributes into our original dataset

#	Name of the State	Alcohol consumption	Heart disease	Fatal alcohol-related accidents
1	Andaman and Nicobar Islands	1.73	20,312	2201
2	Andhra Pradesh	2.05	16,723	29,700
3	Arunachal Pradesh	1.98	13,109	11,251
4	Assam	0.91	8532	211,250
5	Bihar	3.21	12,372	375,000
6	Chhattisgarh	2.03	28,501	183,207
7	Goa	5.79	19,932	307,291



Name of the State.
Liters of alcohol consumed per capita.
Number of fatal heart diseases, measured per 1,000,000 people.
Number of fatal accidents related to alcohol per 1,000,000 people.

Data Transformation

–17, 25, 39, 128, –39 → 0.17, 0.25, 0.39, 1.28, –0.39

Data Reduction

	A1	A2	A3	....	A200
T1					
T2					
T3					
....					
T200					

→

	A1	A2	A3	...	A120
T1					
T2					
T3					
....					
T150					

Data Discretisation

What is Big Data?

- Big Data is a massive collection of data that continues to increase dramatically over time.
- It is a data set that is so huge and complicated that no typical data management technologies can effectively store or process it.
- Big data is similar to regular data, except it is much larger.

Some Examples of Big Data

CHAPTER 1: VALIDATE ASSEMBLED OR OBTAINED BIG DATA SAMPLE

- Establish a sampling strategy for big data testing and identify a representative sample for big data testing.
- Assemble or obtain sample of raw big data according to legislative requirements and organisational policies and procedures.
- Validate big data sample from various sources to ensure that big data is correct.

Big data Or Data validation

It is a process for ensuring data quality and correctness, or it may be defined as data cleaning to guarantee that data is full, unique, and within the specified range.

Data validation is utilised in Extract, Transform, and Load (ETL) processes.

You must transfer data from a database source to a specific data warehouse, where it will be joined with other sets of data for analysis to improve accuracy.

Types of Data Validations

- Data Staging Validation
- Process Validation
- Output Validation

1.1 Establish a sampling strategy for big data testing and identify a representative sample for big data testing

- Big Data Testing
- Testing is the process of ensuring that your software product is of high quality in terms of functionality, performance, user experience, and usability.
- However, when it comes to big data testing, you should concentrate on the applications functional and performance features.
- Any big data application designed to process terabytes of data must have a high level of performance.
- It must be proven that terabytes of data can be processed utilising a commodity cluster with various supporting components.

Strategies behind Sample testing Big Data

- Batch Data Processing Test
- Real-Time Data Processing Test
- Interactive Data Processing Test

1.2 Assemble or obtain sample of raw big data according to legislative requirements and organisational policies and procedures

- Big Data's impact and successful use cases are rapidly increasing.
- Though the potential benefits of Big Data are undeniable, business leaders are concerned.
- Many businesses have successfully implemented Big Data in various functions, and many more are still determining the best way to incorporate it.

7-step process to develop a successful Big Data strategy

CHARLES DARWIN UNIVERSITY AUSTRALIA

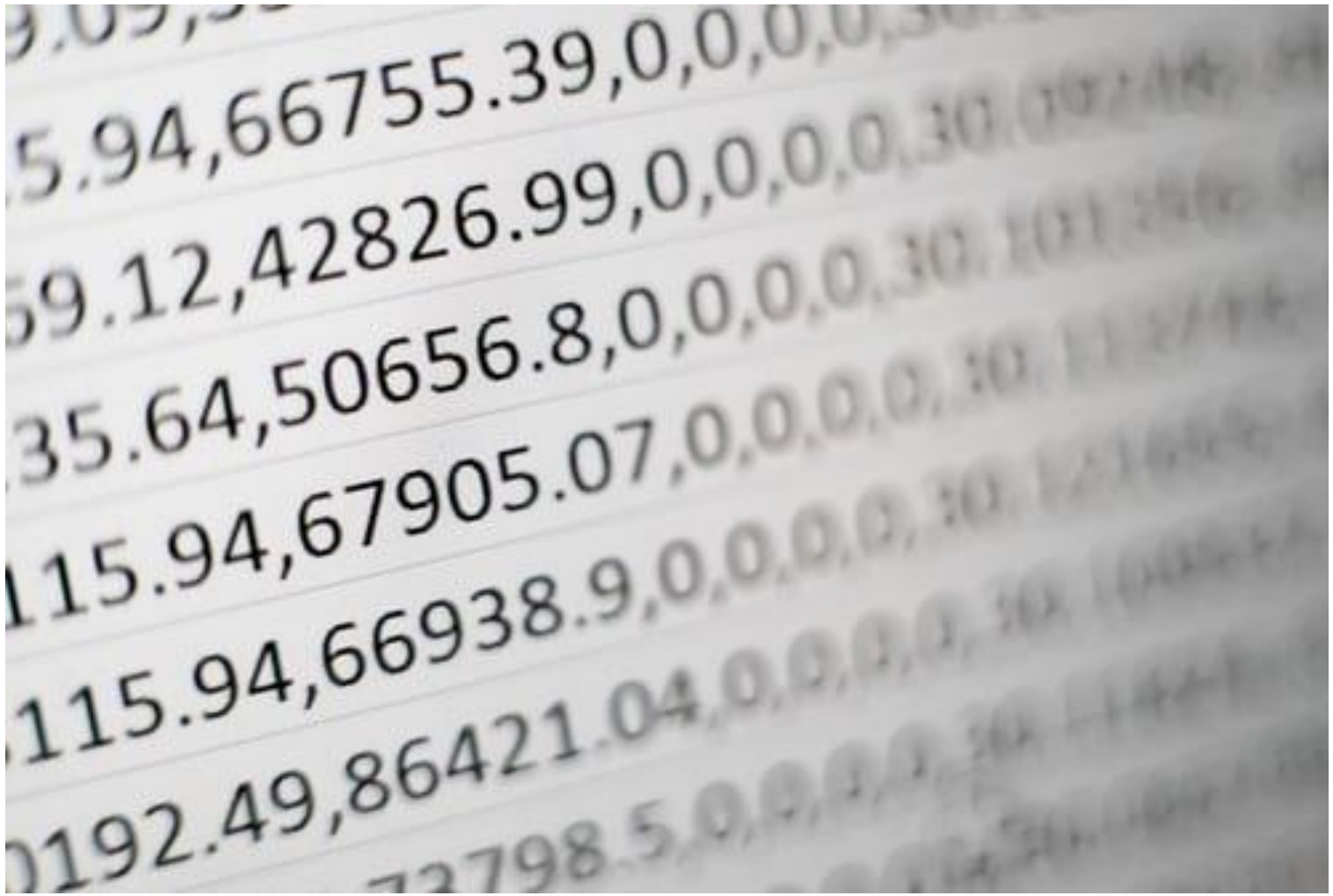
Legislative Requirements for Big Data

Legal discovery

Privacy and Data Protection in a Big Data Context

Policies for big data privacy

- Develop a "Big Data Code of Ethics" and be honest about data utilisation
- Recognise the distinctions between white, black, and grey
- Investigate any biases in the data
- Regulations



1.3 Validate big data sample from various sources to ensure that big data is correct

CHARLES DARWIN UNIVERSITY AUSTRALIA

Steps of Data Validation Process

Determine Data Sample

Database Validation

Data Format Validation

Data Validation Tools for Big Data

Datameer	Talend	Informatica	QuerySurge
ICEDQ	Datagaps ETL Validator	DbFit	Data-Centric Testing

CHAPTER 2 : VALIDATE BIG DATA SAMPLE PROCESS AND BUSINESS LOGIC

- Align datasets to relevant parts of the organisation.
- Implement data aggregation and segregation rules on a small set of sample data and datasets.
- Consult with required personnel to clarify and resolve identified anomalies.
- Conduct performance testing for data throughput, data processing and sub-component performance.

How to perform big data sample testing

- **Data Ingestion**
- **Data Processing**
- **Data Storage**

2.1 Align datasets to relevant parts of the organisation.

- When biases between two datasets are controlled as much as feasible, the overall volume difference between them should be no more than 10% of the preceding data source.
- This is a standard in the business.
- After biases have been considered, your company should set its own standards for what constitutes an acceptable amount of discrepancy across data sources.

Common Reasons Why Analytics Data Deviates From Other Data Sets

- By examining both datasets and adjusting for the area, underperforming servers may be identified.
- Transaction IDs were being filtered out by exclusion filters applied to analytics reporting views.
- Differences in tool time zones
- On order confirmation, the Google Analytics JavaScript took too long to load.
- Many sponsored clicks from IP addresses labelled as "known bots and spiders" by the Interactive Advertising Bureau.

2.2 Implement data aggregation and segregation rules on a small set of sample data and datasets.

What is data aggregation?

- The act of gathering data and presenting it in a summary style is known as data aggregation.
- The data might come from various sources, so merge them into a single summary for data analysis.
- This is an important phase since the volume and quality of data utilised has a big impact on the accuracy of the insights gained from data analysis.
- It is critical to collect high-quality, precise data in sufficient quantities to generate useful findings.
- Data aggregation is beneficial for various purposes, including financial and corporate strategy decisions, as well as product, pricing, operations, and marketing initiatives.

Types of data aggregation

Time aggregation:

- Over a certain period, all data points for a particular resource.

Spatial aggregation:

- Over a certain period, all data points for a collection of resources.

Time intervals for data collection and aggregation

- Reporting period
- Granularity
- Polling period
- Segregation

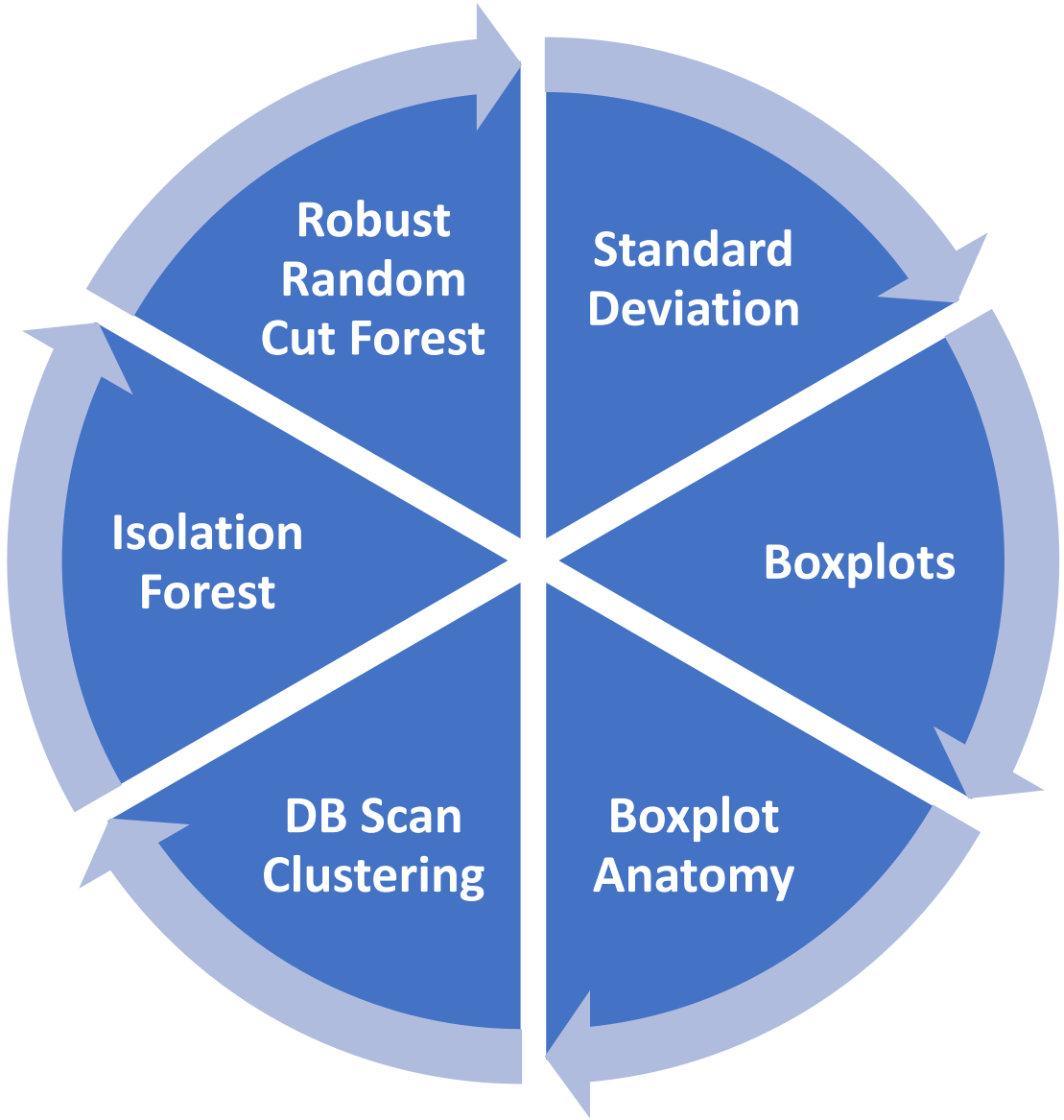
2.3 Consult with required personnel to clarify and resolve identified anomalies

- **Anomaly detection** is a methodology for identifying outliers or odd patterns that do not conform to anticipated behaviour.
- You can consult identified anomalies with your immediate supervisor.
- Anomalies have a wide range of commercial applications, from intrusion detection (finding unusual patterns in network traffic that might indicate a hack) to system health monitoring (finding a malignant tumour in an MRI scan) and credit card fraud detection fault identification in operating systems.

Types of anomalies

- Point anomalies
- Contextual anomalies
- Collective anomalies

Ways to detect Anomalies

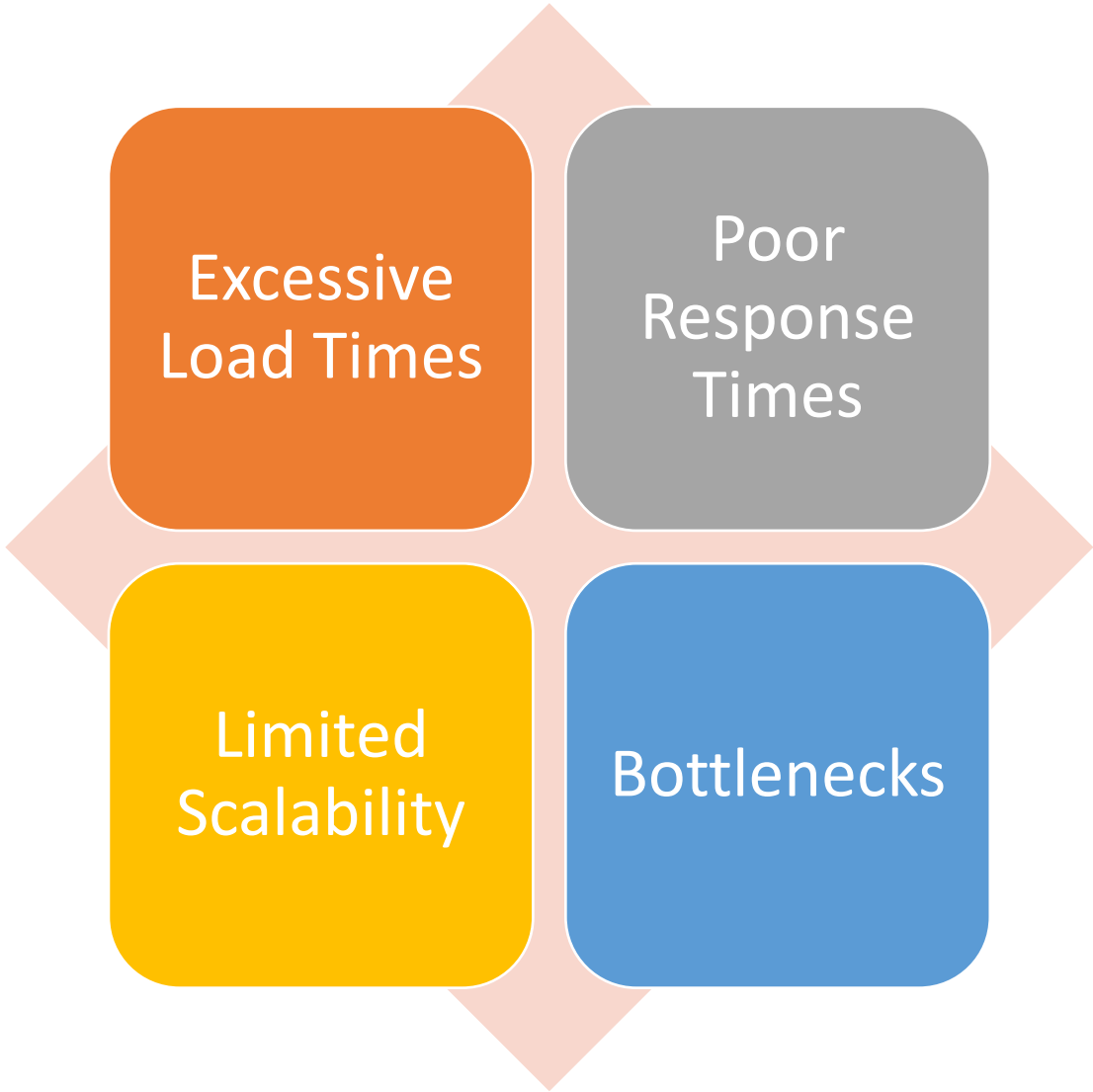


2.4 Conduct performance testing for data throughput, data processing and sub-component performance.

Performance testing may be used to assess various success elements, including reaction times and possible mistakes.

You may confidently detect bottlenecks, defects, and blunders with these performance statistics in hand and determine how to optimise your application to eliminate the problem (s).

What does performance testing measure?



What is the process for performance testing?

What Is Data Processing?

- Any company cannot use data in its basic form.
- The method of gathering raw data and converting it into useable information is known as data processing.
- A team of data scientists and data engineers at a company normally does it in a step-by-step approach.
- The raw data is gathered, filtered, sorted, processed, Analysed, and stored before being displayed in a usable way.
- Data processing is critical for businesses to develop better business strategies and gain a competitive advantage.
- Employees can comprehend and use the data by translating it to an understandable format such as graphs, charts, and texts.

Data Processing Methods

Manual Data Processing

Mechanical Data Processing

Electronic Data Processing

CHAPTER 3: VALIDATE OUTPUT OF CAPTURED BIG DATA SAMPLE AND RECORD RESULTS.

- Design, formulate and select suitable test scenarios and test cases to validate output of big data sample
- Implement selected test scenarios and test cases with big data sample using common testing tools and according to organisational procedures
- Isolate sub-standard data and correct data acquisition paths as required
- Generate and store results of validation activity and associated supporting evidence according to organisational policies and procedures, and legislative requirements

3.1 Design, formulate and select suitable test scenarios and test cases to validate output of big data sample.

Test Case Design

- The way you put up your test cases is referred to as test case design.
- It's critical that your tests are well-designed, or else you risk missing flaws and problems in your product during testing.

Features and formats of common big data sources

- Specification-Based techniques
- Structure-Based techniques
- Experience-Based techniques

3.2 Implement selected test scenarios and test cases with big data sample using common testing tools and according to organisational procedures.

Tools for testing

HDFS (Hadoop Distributed File System)

Hive

HBase

MapReduce

HiveQL

Pig Latin

Challenges faced in Testing Big Data

- Big Data Testing is a difficult procedure that needs the assistance of a highly qualified official.
- The processes for automated Big Data Testing are predetermined and unsuitable for unanticipated mistakes.
- Latency in testing is caused by virtual machine latency, and controlling multimedia is a hassle.
- One of the most difficult aspects of testing is dealing with large amounts of data.
- For many platforms, a test environment and automation should be created.
- Because each component is from a distinct technology, it must be tested separately.
- End-to-end testing is impossible to do with a single tool.
- Designing test scenarios necessitates a high level of scripting.
- To improve performance and test important areas, customised solutions are necessary.

How to Write a Test Script

- Record/playback
- Keyword/data-driven scripting
- Writing Code Using the Programming Language

3.3 Isolate sub-standard data and correct data acquisition paths as required.

- Different data processing architectures for big data have been proposed to address the different characteristics of big data.
- Data acquisition has been understood as the process of gathering, filtering, and cleaning data before the data is put in a data warehouse or any other storage solution.

Key Insights for Big Data Acquisition

- Protocols that allow for the collection of data from any sort of distributed data source (unstructured, semi-structured, structured)
- Frameworks for collecting data from a variety of dispersed sources utilising various protocols.
- Technologies that allow the frameworks to save the data they retrieve in a persistent format.

3.4 Generate and store results of validation activity and associated supporting evidence according to organisational policies and procedures, and legislative requirements

- Validation activities are generated and stored in the Business console and are used to track and manage a test plan for the release and the results.

Validation activity governance

- Change the owner of the activity.
- Change the due date of the activity.
- Change the goals of the activity.
- Assign approvers and testers to the activity

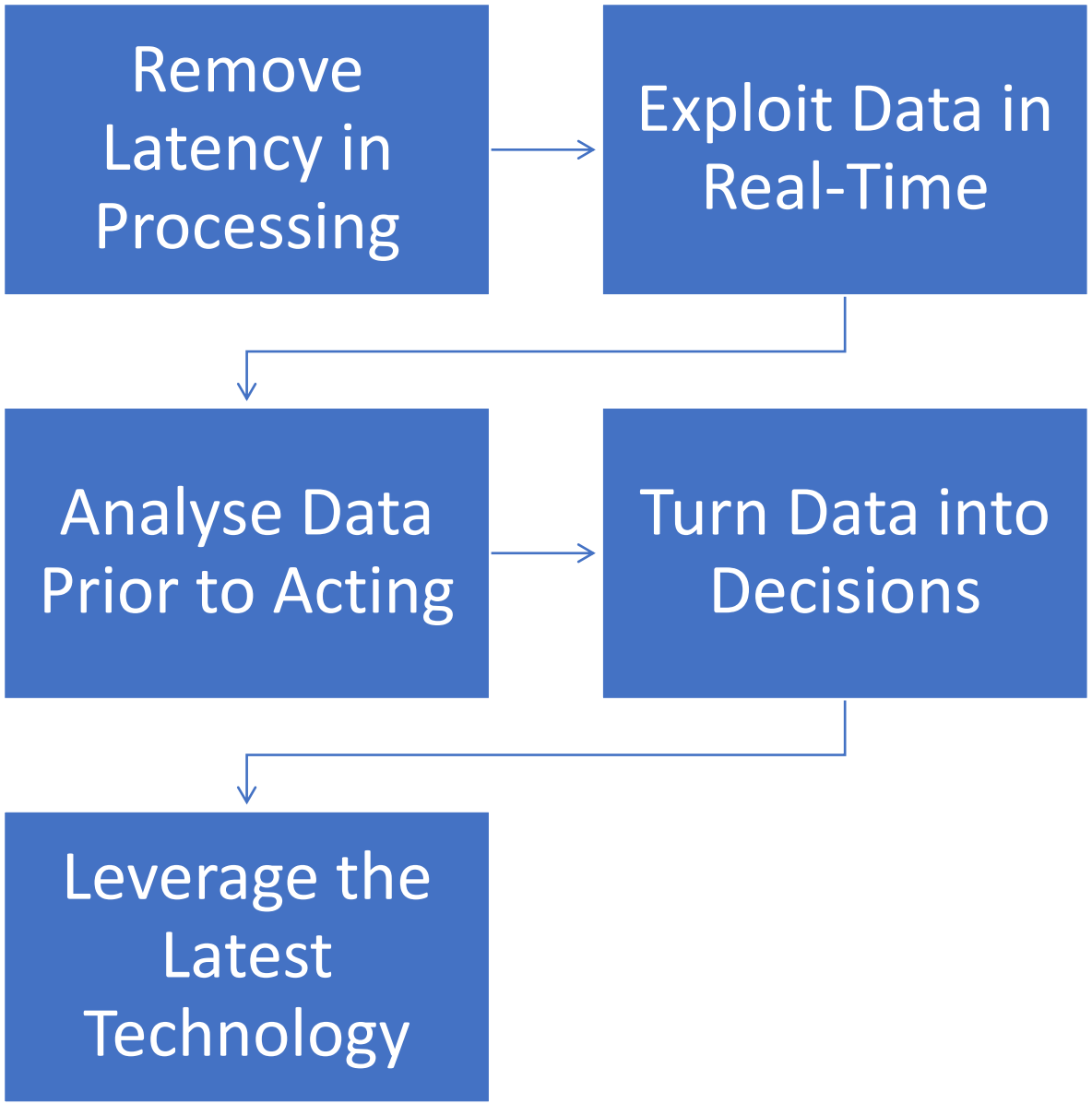
Benefits of storing test results

- Allows storing requirements
- Allows for the storage of test cases as well as their grouping into suites and sets
- Allows for the storage of test results on a per-build basis; a failed test case should be linked to a defect
- Allows for the storage of defect/issue information
- Allows for the storage of build information
- Has code repository (git is preferred)
- Has a Wiki or a similar system for storing knowledge?
- Has integrations with other systems and/or a good API, allowing us to integrate it ourselves

CHAPTER 4: OPTIMISE BIG DATA SAMPLE RESULTS AND DOCUMENTATION

- Perform data cleansing on big data sample following testing according to industry practices and organisational procedures
- Collate validated output of testing, confirming absence of big data corruption in sample
- Recommend configuration optimisation changes based on performance testing results
- Communicate final sample results to required personnel

Optimising the big data



4.1 Perform data cleansing on big data sample following testing according to industry practices and organisational procedures.

- **ETL** is a process that extracts data from several source systems, transforms it (by applying computations, concatenations, and other operations), and then inserts it into a Data Warehouse system.
- ETL stands for Extract, Transform, and Load.

Steps for data cleaning

- Monitor errors
- Standardise your process
- Scrub for duplicate data
- Analyse your data
- Communicate with your team

4.2 Collate validated output of testing, confirming absence of big data corruption in sample.

Big data testing methodologies

- Functional Testing
- Performance Testing
- Data Ingestion Testing
- Data Processing Testing
- Data Storage Testing
- Data Migration Testing

Stages of data processing

- Data collection
- Data preparation
- Data input
- Processing
- Data output/interpretation
- Data storage

4.3 Recommend configuration optimisation changes based on performance testing results.

What is Big Data's role in optimisation?

Big Data is the process of gathering, consolidating, and analysing disparate data from many sources and forms to unearth new insights and create value.

While Big Data originated in the consumer and financial services industries, the global industrial industry has recently been interested in its potential for uncovering important information.

Big Data is, without a doubt, a critical component of Industry 4.0 and the Industrial Internet of Things (IIoT).

Although Big Data can help with process optimisation, it's worth considering why it's just finding its way into the industrial world.

Data, storage, and analytics are three crucial variables to consider.

Some key methods for big data optimisation

- Tune-up’ algorithms
- Remove latency in processing
- Identify and fix errors
- Eliminate unnecessary data
- Bringing it all together

4.4 Communicate final sample results to required personnel

- **Report on evaluation outcomes and obtain sign off from required personnel**
- A report on evaluation outcomes is a written document that explains how the programme was tracked and evaluated.
- It summarises the observations, conclusions, and recommendations from a specific assessment and suggestions on how evaluation outcomes can be used to enhance programmes and make decisions.
- Though evaluation is a continuous process, the word "final" refers to the final report of a funding cycle or a particular evaluation task.

How do you write an evaluation report?

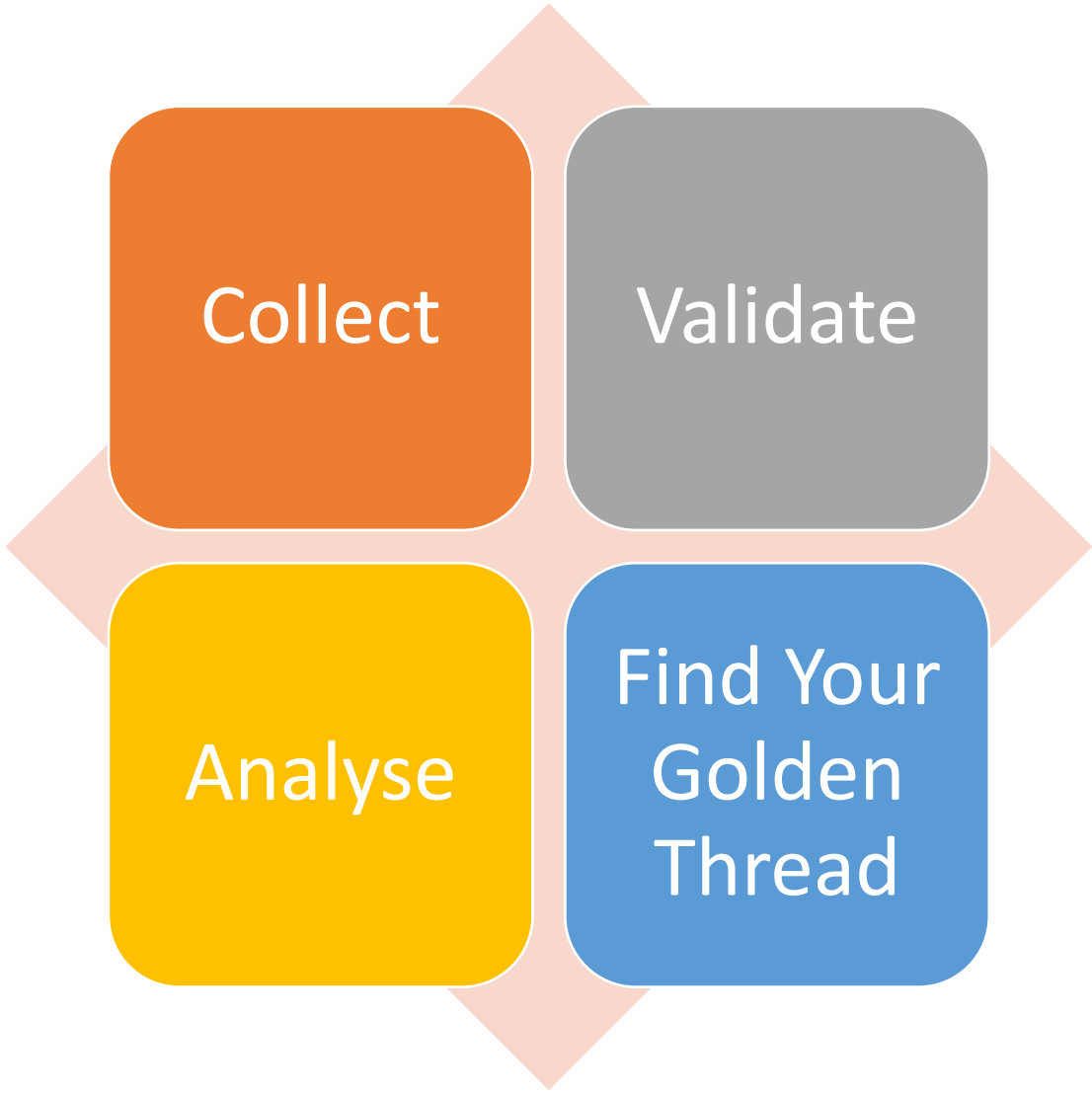
- Intended use and users
- Program description
- Evaluation focus
- Data sources and methods
- Results, conclusions, and interpretation
- Use, dissemination, and sharing

Methods to communicate final sample results

- Meeting
- Video conference
- Telephone conference
- E-mail
- Report
- Presentation
- Forum

4.5 Industry protocols and procedures required to write queries and scripts for big data testing

Protocols and Procedure for big data to write scripts and queries



Big Data Testing stages

1

Step 1: Data Staging Validation

2

Step 2: “Map Reduce” Validation

3

Step 3: Output Validation Phase

Any Questions?

CHARLES DARWIN UNIVERSITY AUSTRALIA

What will I learn?

In this chapter, you will learn about the following:

1. Establish a sampling strategy for big data testing and identify a representative sample for big data testing.
2. Assemble or obtain sample of raw big data according to legislative requirements and organisational policies and procedures.
3. Validate big data sample from various sources to ensure that big data is correct.

Big Data Testing

Testing is the process of ensuring that your software product is of high quality in terms of functionality, performance, user experience, and usability. However, when it comes to big data testing, you should concentrate on the applications functional and performance features. Any big data application designed to process terabytes of data must have a high level of performance. It must be proven that terabytes of data can be processed utilising a commodity cluster with various supporting components. Processing should be quicker and more precise, which necessitates extensive testing.

Features of big data

- Predictive analysis is one of the most significant benefits of Big Data. Big Data analytics technologies can reliably forecast outcomes, helping businesses and organisations to make better decisions while also improving operating efficiencies and lowering risks.
- Businesses all around the globe are simplifying their digital marketing tactics to improve the entire consumer experience by using data from social media platforms utilising Big Data analytics technologies. Big data helps businesses get insight into their customers' pain areas and enhance their products and services.
- Big Data is accurate because it integrates relevant data from various sources to provide highly actionable insights. Almost 43% of businesses don't have the tools they need to filter out unnecessary data, which ends up costing them millions of dollars to sort through. Big Data technologies can help you save time and money by reducing this.
- Big Data analytics might aid businesses ingenerating more sales leads, increasing income. Big Data analytics technologies are being used by businesses to determine how well their products/services are performing in the market and how they react to them. As a result, they will better understand where they should put their time and money.
- You can always keep one step ahead of your competitors using Big Data insights. You may do a market analysis to determine what types of promos and offers your competitors are offering and then develop better deals for your clients. Furthermore, Big Data insights enable you to discover client behaviour to better understand client patterns and give them amore "personalised" experience.

Strategies behind Sample testing Big Data

Testing an application that manages terabytes of data would need new levels of competence and creative thinking. Three scenarios serve as the foundation for the Quality Assurance Team's core and significant tests. Specifically,

- Batch Data Processing Test
- Real-Time Data Processing Test
- Interactive Data Processing Test

Batch Data Processing Test:

The Batch Data Processing Test consists of test processes that execute data when applications are handled in Batch Processing mode utilising Batch Processing Storage Units such as HDFS. Batch Process Testing is primarily concerned with

- putting the program through its paces with erroneous inputs
- altering the amount of data collected

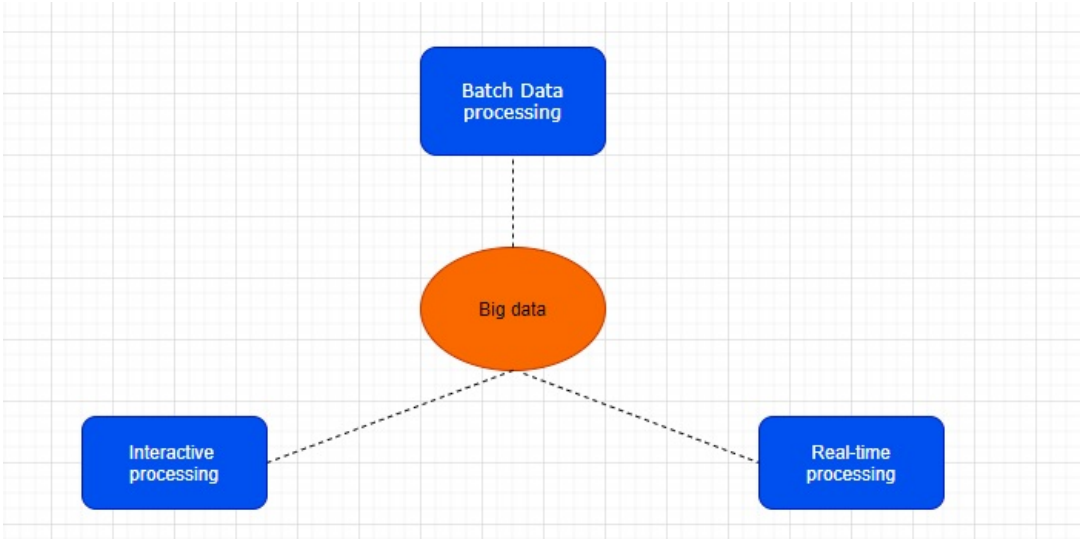
Real-Time Data Processing Test:

When the program is in Real-Time Data Processing mode, the Real-Time Data Processing Test interacts with the data. Real-Time Processing technologies such as Spark are used to operate the program.

Real-time testing entails running a program in a real-time environment and checking for stability.

Interactive Data Processing Test:

The Interactive Data Processing Test combines real-world test protocols with the application to simulate how a real-world user might interact. HiveSQL and other Interactive Processing technologies are used in the Interactive Data Processing mode.



We must test the volume and variety of data taken from various databases and loaded as well as processed on a Data Warehouse or Hadoop System while testing Big Data. This testing is classified as functional testing.

Performance Testing necessitated testing the Velocity of Data retrieved from multiple databases and uploaded to the Hadoop System.

As a result, as a plan or strategy, we should focus on Functional and Performance Testing of Big Data Testing.

In Big Data Testing, the tester must ensure that a large volume of data is processed utilising commodity hardware and related components. As a result, data quality plays a significant role in Big Data testing. Verifying and validating the quality of data is critical.

Testing transactional and non-transactional sources of big data

Types of data sample we used

- Transactional
- Non-transactional

Transactional Data:

Transactional data is information about an organisation's transactions, such as information collected when sold or purchased. Customer, product, and supplier data are examples of master data that are used in various transactions. In most cases, master data does not change and does not require creation with each transaction.

Examples:

- Purchases: purchases by a customer.
- Invoices: A bill for products.
- Debits: Cash or fund added to accounts for, e.g. Ticket cancellation refund etc.
- Credits: cash or fund remove from the account for, e.g. withdrawing money from ATM’s or banks

Non Transactional data:

Non-transactional is a term that refers to something that isn't (Unlike Transactional Data, this information is useful to the organisation for a longer period.) To use a transaction as an example; Consider that you have a database that deals with client orders, payments, and other billing information; the data is critical.

Examples:

- Customer: Name, Preferences
- Product: Name, Hierarchy
- Site/Location: Addresses
- Account: Contracts Detail

Organisational procedures to sample raw big data

Big Data's impact and successful use cases are rapidly increasing. Though the potential benefits of Big Data are undeniable, business leaders are concerned. Many businesses have successfully implemented Big Data in various functions, and many more are still determining the best way to incorporate it.

We have developed a 7-step process to assist you in developing a successful Big Data strategy. Let's take a look at the steps you'll need to take to strategically integrate Big Data into your current business activities:

Identify What You Want: Your end goal most influences the shape of your overall strategy. You must decide whether you want to increase customer service rep efficiency, improve operational efficiency, increase revenue, provide a better customer experience, or improve marketing. Your objective should be precise, certain, and direct. Any strategy that is solely focused on exploring possibilities is doomed to fail. You can choose a methodology, hire employees, and select the appropriate data sources based on your goal.

Leverage a Proven Big Data Strategy: There are four tried-and-true methods for developing a viable Big Data strategy. Depending on your end goal and the availability of data, you can choose one of the following big data strategies to achieve successful results:

- Performance Management: It entails making decisions about storemanagement and operational supremacy based on transactional data such ascustomer purchase history, turnover, and inventory levels.
- Data Exploration: This approach heavily relies on data mining andresearch to uncover solutions and correlations that would be difficult todiscover with in-house data.
- Social Analytics: Non-transactional data on social media and reviewsites such as Facebook, Twitter, and Google+ is measured using socialanalytics. It is based on an examination of the conversations and reviews thatoccur on these platforms. It reveals three key analytics: awareness,engagement, and word-of-mouth.
- Decision Science: Non-transactional data, such as user-generatedcontent, ideas, and reviews, are used in decision science experiments andanalyses. Decision science is more concerned with the exploration ofpossibilities than with the measurement of known objectives.

Identify Infrastructural Changes: To leverage Big Data, particularly historical databases, your company may need to make numerous infrastructure changes. If old company data was stored in traditional formats, it might be difficult to run complex algorithms and analyses.

Establish Talent Pool: One of the most important aspects of developing a Big Data strategy is human resources. Your Big Data team must include statisticians to interpret data, business analysts to communicate insights to decision-makers and key decision-makers who can lead the team.

Obsess Over Customer Satisfaction: The primary application of Big Data is to generate insights that can help businesses better serve their customers. Customer-focused marketing is the new way to approach the market and generate revenue.

Ensure Usability: It is not uncommon for statisticians' insights to be beyond the comprehension of staff. The analysts' data, analytics, and insights must be precisely communicated to the implementation team.

Be Agile: This goes without saying. While implementing disruptive technologies, many hurdles might come up that no one initially thought about.

Legislative Requirements for Big Data

Organisations that utilise Big Data must additionally comply with any industry- or region-specific legislation, such as the Health Insurance Portability and Accountability Act (HIPAA), which oversees the use of "Protected Health Information" and the Children's Online Privacy Protection Act (COPPA).

Legal discovery

Businesses must be aware that using big data analytics may expose them to legal discovery from rival plaintiffs and government regulators. With the advent of big data, technical and other limits for data retrieval have diminished, and firms may be required to furnish the raw data that underpins their big data studies. This might include sensitive proprietary information and personally identifiable information (PII). A corporation must do a legal risk analysis of the information, players, and concerns involved before publishing a big data study.

Once the legal discovery process has started, it may be difficult for a corporation and its lawyers to limit the scope of the inquiry, resulting in the production of more material than is required. The legal hurdles to such broad inquiries are still evolving, and there are no well-established industry best practices in place.

Privacy and Data Protection in a Big Data Context

This section, which examines some of the pertinent difficulties and possibilities linked to privacy and data protection, aims to demonstrate some of the complexities that particular ideas, principles, and duties may bring in the context of a disruptive technology like big data.

Principles:

The GDPR provides some data protection standards that must be followed when processing personal data, most of which are being tested by big data's core characteristics.

- According tothe concept of " lawfulness, " each processing of personal datashould be based on a legal foundation, according to the concept of"lawfulness" (see next section).
- Unless theindividual already possesses this information, the principle of "fairnessand transparency" requires the controller to offer information to personsabout its processing of personal data. The transparency principle, which statesthat “individuals must be given clear information on what data is processed,including data observed or inferred about them; better informed on how and forwhat purposes their information is used, including the logic used inalgorithms,” can be particularly challenging in a big data context, where thecomplexity of the analytics renders the processing opaque.
- According tothe " purpose restriction " concept, personal data must be gatheredand processed for specific, stated, and legal reasons, according to the"purpose restriction" concept. To begin with, every processing ofpersonal data must have a clearly stated purpose to be allowed. This may beespecially challenging in the case of big data since "it may still beunknown at the moment personal data is acquired for what purpose it willsubsequently be utilised." However, the broad assertion that the data isbeing gathered for (any hypothetical) big data analytics is not specificenough.
- According tothe " data minimisation " concept, personal data must be sufficient,relevant, and restricted to what is necessary for connection to the purposesfor which they are processed, according to the "data minimisation"concept. The terms "data minimisation" and "big data"appear to be mutually exclusive at first glance. Indeed, "the perceivedpotential in big data gives incentives to gather as much data as possible andto preserve this data for as long as possible for still undefined futureapplications," according to the report.
- Personal datamust be stored in a form that allows data subjects to be identified for nolonger than is required for the purposes for which they are processed,according to the "storage limitation" principle. Given that dataretention durations are always context-specific, the GDPR does not prescribethem. Big data analytics is an excellent example of the advantages of processingpersonal data for a longer period, as well as the challenges that may occur dueto the storage restriction principle.

Privacy Law for Big data

Both federal and state legislation govern data protection in Australia. 1st In this post, we'll look at the Privacy Act, 1988 (Cth), Australia's federal privacy law. The Privacy Act is based on 13 listed data collection and use standards (the "Australian Privacy Principles" or "APP"). The Office of the Australian Information Commissioner ("OAIC") is in charge of enforcing the Privacy Act and dealing with complaints.

Data processing and other obligations of APP entities

Individuals, commercial sector businesses with an annual turnover of more than AUD 3 million, and all Commonwealth Government and Australian Capital Territory government agencies are all covered by the Privacy Act.

The Privacy Act, through the APPs, places restrictions on how APP organisations acquire, store, utilise, and disclose personal information. For example, other than sensitive information, personal information must only be gathered if it is "reasonably required. “As a result, no consent is necessary for the acquisition of personal data. However, collecting sensitive information also necessitates the individual's agreement, legal authorisation/requirement, or a court/tribunal order.

APP entities must notify persons about this collection at the time of collection or as soon as practical afterwards. In some cases, such as disclosing sensitive information about an individual for direct marketing or using data gathered for a secondary purpose, consent is necessary.

Personal information must be protected from abuse, interference, and loss and unauthorised access, alteration, and disclosure by APP entities.

When APP organisations have reasonable grounds to suspect that an "eligible data breach" has occurred, they must inform the OAIC. [Number 16] Unauthorised access to or disclosure of personal information that is likely to cause “serious harm” to the individual is considered a data breach. (17) as far as is practical, the APP entity must also tell the individual in question while also advising them on how to respond to the breach. [Nineteen] These amendments to the Privacy Act were made by an amendment in 2017 and resulted in a 712% rise in data breach notifications in a year (compared to the voluntary notification regime prior to that)

APP entities shall take reasonable effort to put in place methods, processes, and systems that will assure APP compliance and handle complaints. This is referred to as a "privacy management strategy" by the OAIC. It demands companies to use a privacy-by-design approach as part of its template. Conducting privacy impact assessments is part of this.

Policies for big data privacy

Develop a "Big Data Code of Ethics" and be honest about data utilisation:

Our marketing team proactively solicited everyone's feedback and produced "Evolv's Workforce Big Data Code of Ethics" as these tales began to emerge. It wasn't designed to be a formal document, but rather a statement of our beliefs and attitude about what data was gathered, how it was managed, and what data was on- and off-limits when it came to making job recommendations. It was an excellent chance to conduct some corporate soul-searching and determine our values. More significantly, we now had a document to refer to and give to anyone who called to inquire about our data policy.

Recognise the distinctions between white, black, and grey:

Employees felt comfortable expressing when they thought we were swimming into muddy waters because of the basis provided by the "Big Data Code of Ethics." For example, we were looking for any characteristics of job applicants that would predict their job success when we famously discovered that job applicants who logged into our platform using Chrome and Firefox browsers stayed on the job longer and performed better than those who logged on using the Internet Explorer or Safari. We could have utilised this data to fine-tune our algorithms and make the evaluation more accurate, but we quickly realised that it raised some ethical concerns. Not only did we feel that browser usage was significantly linked to one's age, but we also believed that scraping information about one's browser fell under the category of "creepy data acquisition." We pondered how people would respond if they knew. We ultimately decided not to include it in our grading process since we believed it was one of those ambiguous areas that should be avoided.

Investigate any biases in the data:

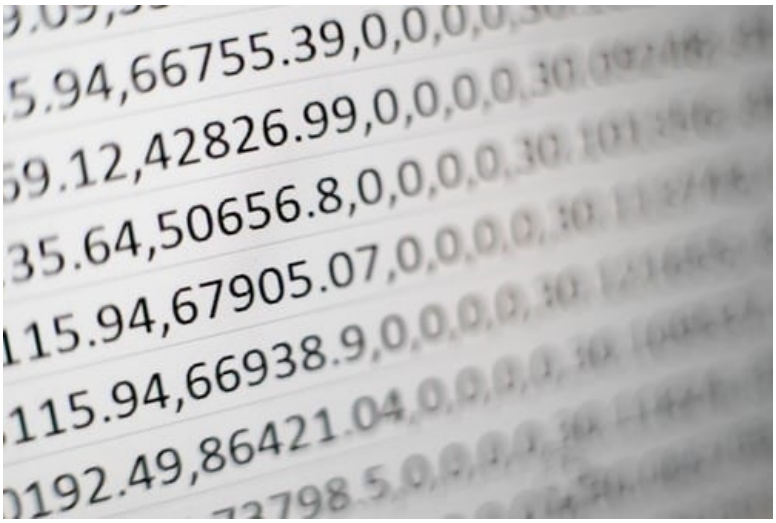
We realised this was a sensitive area where people had a collective sense of justice since we were utilising data to make employment recommendations. There are laws in place to guarantee that protected classes are not discriminated against.

Regulations:

The scope of this does not allow for a more in-depth examination of the findings. However, the perspectives voiced by research participants and represented above give some insight into the diversity of perspectives on Big Data. Greater knowledge of all sections of society's perspectives on Big Data can aid in the formulation of suitable regulations, especially when dealing with sensitive issues like the use of Big Data to support national security and law enforcement.

In the sections following, we'll look at how to deal with this problem. We argue that traditional legal instruments such as constitutional principles, legislation, regulations, and case law are insufficient to regulate Big Data in the commercial and public sectors.

To be productive in the Web of Data, various semantic and algorithmic tools are becoming increasingly important.



Steps of Data Validation Process

1. Determine Data Sample

If you need to validate a huge quantity of data, you'll need a sample rather than the entire dataset. To ensure the project's success, you must first determine the volume of the data sample and the error rate.

2. Database Validation

You must confirm that all requirements are met using the current database throughout the database validation procedure. To compare source and destination data fields, unique IDs and the number of records must be determined.

3. Data Format Validation

Determine the data's overall capabilities and the variation that the targeted validation demands, and then look for inconsistencies, duplicate data, null field values, and inappropriate formats.

Data Validation Tool for Big Data

There are a number of tools that can give the best data validation process with the best performance and acceptable outputs, including:

- Datameer
- Talend
- Informatica
- QuerySurge
- ICEDQ
- Datagaps ETL Validator
- DbFit
- Data-Centric Testing

Challenges come in data validation

Validating data can be difficult for a variety of reasons:

Because data may be spread across different databases in your business, validating the database might be difficult. It's possible that the data is segregated or out of date.

Validating the data format might take a long time, especially if you have huge datasets and want to do it manually. However, sampling the data for validation can assist to cut down on the amount of time required.

Data validation and ETL

Validating data, whether manually or by scripting, can take a long time. After you've validated your data, though, a contemporary ETL tool like Alooma may help you speed up the process. You can identify which faults can be rectified at the source and which faults an ETL tool can repair while the data is in the pipeline as part of your data review. As data is transported to your data warehouse, you may automatically integrate, clean, and transform it.

While data validation might be difficult, Validation can help you automate it. Our data professionals can assist you in planning, executing, and maintaining your data pipeline once you've decided what data you want to verify and transport.

What will I learn?

In this chapter, you will learn about the following:

1. Establish a sampling strategy for big data testing and identify a representative sample for big data testing.
2. Assemble or obtain sample of raw big data according to legislative requirements and organisational policies and procedures.
3. Validate big data sample from various sources to ensure that big data is correct.

Assumptions of Digital Dataset

When biases between two datasets are controlled as much as feasible, the overall volume difference between them should be no more than 10% of the preceding data source. This is a standard in the business. After biases have been considered, your company should set its own standards for what constitutes an acceptable amount of discrepancy across data sources.

There is knowledge of the previously existing data set. The definitions, biases, and measurement techniques for the dataset are well-known.

The previously existing dataset has measured the relevant metrics consistently. The dataset cannot measure things inconsistently. For example, if your dataset cannot track mobile phone users, that’s ok because the bias is consistent; however, if the dataset has had portions deleted at the CRM team’s discretion without documentation or review of reasoning, then normalisation should not be done

Steps to Align Dataset

- Identify the metric(s) that will be used to normalise the dataset. There should be no more than 1-2 relevant measures, and they should be the most relevant metrics for the organisation's performance.
- Inquire with the owner(s) of the internal dataset whether there are any existing biases between the internal dataset and the analytics tool. Consider the following scenario: This dataset includes orders from the contact centre, which are not included in the analytics tool.
- Request that the internal dataset owner(s)supply a few days' worths of data that has been adjusted for the dataset biases mentioned in #2. The duration of time and the size of the dataset should be adequate to record at least a few hundred instances of the KPI.
- Compare the topline volume of the KPI captured in internal data vs. the analytics tool over time.
- You're done and ready to go on to the next dataset if the difference between the two datasets is within an acceptable range (be sure to document the details on project closure).
- Analyse the data inside the analytics tool to uncover connections and probable reasons if the difference between the two datasets is outside the permissible range. For instance, this dataset is missing many transactions from the morning, or Leads from Apple-branded devices are absent from this dataset. Continue to step 5.
- Provide the results to your stakeholder and ask for their help in determining the source of the discrepancy between the datasets. Rinse and repeat the processes until all relevant KPIs have been met.

Common Reasons Why Analytics Data Deviates From Other Data Sets

The trickiest aspect of this approach is establishing the source of variation between datasets. Through innovative investigation across various customer accounts, we’ve discovered:

- By examining both datasets and adjusting for the area, underperforming servers may be identified.
- Transaction IDs were being filtered out by exclusion filters applied to analytics reporting views.
- Differences in tool time zones
- On order confirmation, the Google Analytics JavaScript took too long to load.
- Many sponsored clicks from IP addresseslabelled as "known bots and spiders" by the Interactive Advertising Bureau.

The route to attaining corporate goals is lit with compelling insights when an effective firm data strategy that is tightly connected with business strategy has been deliberately designed. Improved data quality implies greater consumer insights (better customer engagement and retention), clear priorities (better marketing ROI), and a clear picture of cause and effect across all departments due to these efforts.

What happens when you don’t align data strategy to support business strategy?

Consider it this way: your company's data strategy is its backbone. Everything built on top of the foundation will be jeopardised if the foundation is cracked or shattered into pieces.

- All business units are making poor decisions.
- Customer retention is poor due to ineffective marketing.
- Inability to foresee properly Due to a lackof competitive knowledge
- Revenue targets were not met.
- Unnecessary expenditures have accumulated throughout the organisation.

The present data system may not provide correct insights, or accessing the data required may be difficult. Individuals accessing and maintaining duplicate data, as well as redundant applications, increase expenses. Most significantly, developing a single accurate perspective of customers and revenue may be difficult without a unified system for delivering business-aligned insights.

Aligning across the business

Working closely with business executives to obtain knowledge of the company's main problems is key to achieving alignment. A solid data strategy will first and foremost assist in resolving those difficulties through a system of technical infrastructure and clearly defined procedures.

This degree of clarity will aid in the development of a targeted plan that utilises just the facts that the organisation need. It's simple to boil the ocean if business results aren't considered from the outset.

Building good working connections with corporate executives will be a side benefit of this degree of alignment. Creating a new data strategy is not a quick (or inexpensive) task, and their success depends on it.

Begin by defining the desired state, then identify gaps

Many businesses today have a disorganised and reflexive approach to data. It's frequently stored, accessed, and used differently by different departments inside the firm, resulting in inconsistencies and inaccuracies. Data isn't living up to its full potential in this setting.

Within a firm, data should be used to support and ultimately push forward business goals. Only by adopting a strategy for gathering, organising, and using data that supports the company's goals rather than the particular goals of each team can this be accomplished. A solid data strategy isn't, and shouldn't be, self-contained.

To emphasise a crucial point, the only method to get the firm to the desired condition is to collaborate with other business executives to grasp the organisation's short and long-term objectives. What do you want the business to look like in six months? Is it a year? How about five years? What role does data play in achieving those objectives? The idea is to go beyond what data can now accomplish for a company and instead aim to embrace what data may accomplish for a company to help them achieve their business objectives more quickly. This is where you provide the desired data state.

Once the intended state has been established, it's time to examine the present data environment in light of the newly specified business objectives. Identify all existing data sources, how they're stored and accessible, and how data insights are created and used in this study. This exercise aims to find the holes in the present data "strategy" about the intended state. This will assist in determining the adjustments that need to be made in terms of technology and procedure when establishing the new data strategy.

Remain agile

We live in a world that is becoming increasingly dynamic. When a result, it's critical to be fluid and opportunistic, iterating on the data strategy as new information about customer preferences and/or company needs emerge.

For e.g.:

The marketing aims of budget hotel business Red Roof Inn are to be flexible and contextual to better connect and interact with customers around good experiences. The hotel business understood that bad weather might help them increase last-minute reservations from stranded passengers, but only if they could find out how to take advantage of dynamic weather variations. To meet their objectives, they coordinated their data approach to assist detect and advise when severe weather may cause flights at adjacent Red Roof Inn locations to be cancelled.

With an estimated 2% to 3% of flights being cancelled every day, this translates to around 500 planes not taking off and 90,000 people being stranded. Red Roof Inn utilises this information to target stranded clients through mobile advertising campaigns that include tailored messaging like "Stranded at O'Hare?" to encourage digital reservations. Look into the Red Roof Inn.' The advertising sent a timely and relevant offer to each receiver, appearing at precisely the proper micro-moment in time.

Alignment techniques work across all five capabilities referenced above. At a high level, alignment is accomplished via the following steps:

Understand business strategy: What makes a difference if all analytics expectations are met? How does the financial picture look?

Decompose strategy into how data will be used to help meet strategic goals

We frequently give what constituent requests rather than what the firm requires. Business needs, not wishes, must be articulated in strategies.

Decompose data usage into the critical parts of data element, metrics, dimensions, lists and values

Every “requirement” for managing data can be stated as a metric, a fact, or a contributor to that metric or fact. As a result, the demand is kept in a business context.

Understand patterns of use

Some data applications, such as real-time analytics, may necessitate large volumes of data and high velocity. Others, such as a customer happiness score, may require pre-calculated and consistent data yet drives interactions.

What is data aggregation?

The act of gathering data and presenting it in a summary style is known as data aggregation. The data might come from various sources, so merge them into a single summary for data analysis. This is an important phase since the volume and quality of data utilised has a big impact on the accuracy of the insights gained from data analysis. It is critical to collect high-quality, precise data in sufficient quantities to generate useful findings. Data aggregation is beneficial for various purposes, including financial and corporate strategy decisions, as well as product, pricing, operations, and marketing initiatives.

Raw data, for example, can be aggregated over some time to provide statistics like average, minimum, maximum, total, and count. You can evaluate the aggregated data to get insights about specific resources or resource groups once the data has been aggregated and put to a view or report.

There are two types of data aggregation:

- Time aggregation - Over a certain period, all data points for a particular resource
- Spatial aggregation - over a certain period, all data points for a collection of resources

Time intervals for data collection and aggregation

Data is collected and presented in a view or report within the context of various time intervals:

Reporting period:

The period during which data is gathered for the presentation. A resource summary table, for example, may comprise data collected for a specific network device over a single day. A reporting period might comprise both aggregated and raw data elements (data that has not been aggregated). Daily, Weekly, Monthly, Quarterly, and yearly reporting periods are all supported.

Granularity:

Data points for a certain resource or collection of resources are collected for aggregate over a specific period. For example, if you wish to determine the average of data points gathered during 5 minutes for a certain resource, the granularity is 5 minutes. Depending on the reporting period and view or report type, granularity might range from one minute to one month. Data View dynamically aggregates data at a resolution of less than a day. For higher granularity values, Data Channel aggregates data.

Polling period:

The length of time controls how frequently data is sampled from resources. A collection of resources, for example, may be polled every 5 minutes, resulting in a data point for each resource every 5 minutes: the polling duration and granularity influence a geographical aggregation's outcome. For example, assume you wish to calculate the average of a collection of data points collected during 10 minutes for a collection of devices (the granularity). The result is the average of single data points obtained for each device if the polling duration is 10 minutes. However, if the polling time is five minutes, each device gets sampled twice within the granularity period of ten minutes.

Aggregation is most common in patchy environments2 where animals congregate around supplies (e.g. food sources or shelters during resting period4). In most cases, the procedure is quite quick, and the time spent outside the patches by people may be deemed minimal in terms of dynamics5. When there are numerous patches, unequal distribution of the population among patches (from now on referred to as aggregation) can emerge from inter-patch heterogeneities as well as social interaction between people in the event of a collection of identical patches.

Segregation:

Segregation can happen for a variety of reasons, including distinct environmental preferences or agonistic interactions between species. Environmental restrictions, particularly the carrying capacity of patches, have been shown in several studies to play a role in the ensuing segregation patterns5.

Asymmetrical encounters can be agonistic, neutral, or attractive in addition to being agonistic, neutral, or attractive. Chemical communications investigations, for example, reveal that distinct species can exchange components (kairomones) to attract one other, but that in other circumstances, just one species attracts the other and not the other way around23,28. Many factors might be at play in

the patterns revealed in these researches. Still, no systematic relation has been shown between individual-level variation and the main features of collective behaviour.

The number of distinct species, environmental restrictions, and heterospecific interactions all influence various aggregation-segregation patterns without requiring individuals to modify (modulate) their behavioural algorithm. On the one hand, we develop a generic model that incorporates the general laws of inter attraction between conspecific and heterospecific contacts (in the absence of agonistic behaviour) and environmental variables (sub-group composition and patch carrying capacity) on the other6. We demonstrate how transitions between distinct patterns can occur in an ecosystem made up of similar patches whose selection cannot be influenced.

We demonstrate how transitions between distinct patterns can occur in an environment made up of similar patches whose selection cannot be based on heterospecific preferences. We generate the bifurcation diagram of the entire set of steady-state solutions by deriving a set of coupled differential equations for the model variables. The research is supplemented with a stochastic description based on Monte Carlo simulations that take variations into account.

Segregation is happening as a consequence of the limited size of the carrying capacity. It is worth noting that the model analysed and the results obtained are largely independent of the particular types of signals or cues used by different species. Indeed, the model can be viewed as a network of feedback between individuals of the same sub-group and of different sub-groups whose presence or absence and strength determine the observed behaviours. Without denying the importance of the specificities of each of the cases encountered in nature, the model is therefore generic and applicable to a whole spectrum of situations involving different feedbacks.

Even in the absence of aggressive agonistic behaviour, segregate across patches and for transition between various patterns. The model may be thought of as a feedback network unaffected by the signals or cues used in mixed group interactions. As a result, its predictions apply to a wide range of scenarios, including social insects, and it sheds light on the mechanisms at work.

Anomaly detection is a methodology for identifying outliers or odd patterns that do not conform to anticipated behaviour. You can consult identified anomalies with your immediate supervisor. Anomalies have a wide range of commercial applications, from intrusion detection (finding unusual patterns in network traffic that might indicate a hack) to system health monitoring (finding a malignant tumour in an MRI scan) and credit card fraud detection fault identification in operating systems.

What Are Anomalies?

Before we begin, it's vital to set some ground rules for what constitutes an abnormality. Anomalies are classified as follows:

Point anomalies:

If one piece of data differs significantly from the others, it is considered anomalous. Detecting credit card fraud based on "amount spent" is a business use case.

Contextual anomalies:

The anomaly is context-dependent. In time-series data, this form of anomaly is prevalent. Business use case: It's customary to spend $100 on food every day during the holidays, but it's unusual otherwise.

Collective anomalies

A group of data examples together can aid in the detection of abnormalities. Business use case: Someone tries to copy data from a distant workstation to a local host without permission, which would be identified as a possible cyber assault.

Noise reduction and novelty detection are comparable but not identical to anomaly detection. Novelty detection is concerned with detecting an undetected trend in fresh observations that aren't part of the training data, such as a sudden interest in a new YouTube channel around the holidays. The practice of immunising analysis from the occurrence of undesirable data or eliminating noise from an otherwise useful signal is known as noise removal (NR).

Anomaly Clarify Techniques

The most basic method for detecting data abnormalities is to mark data points that differ from typical statistical features of a distribution, such as mean, median, mode, and quantiles. Assume that an anomalous data point deviates from the mean by a particular standard deviation. Because time-series data isn't static, traversing means over it isn't easy. To compute the average over the data points, you'd need a rolling window. This is known as a rolling average or moving average in technical terms, and it's used to smooth short-term changes while highlighting long-term ones.

Challenges

The low pass filter may be used to detect abnormalities in basic use cases. However, there are some scenarios when it does not function. Here are a few examples:

- Because the line between normal and abnormal conduct is often blurry, the data contains noise that might be mistaken for aberrant activity.
- As hostile enemies continuously adapt, the concept of abnormal or normal may alter regularly. As a result, the moving average barrier may not always be applicable.
- Seasonality is the basis for the pattern. This necessitates more advanced techniques, such as dissecting the data into different trends to determine the shift in seasonality.

Ways to detect Anomalies

Standard Deviation:

According to statistics, if a data distribution is roughly normal, around 68% of the data values are within one standard deviation of the mean, 95% are within two standard deviations, and 99.7% are within three standard deviations.

As a result, any data point with a standard deviation of more than 3 is likely to be abnormal or outlier.

```python
import numpy as np
import matplotlib.pyplot as plt
seed(1)


# multiply and add by random numbers to get some real values
data = np.random.randn(50000)  * 20 + 20

# Function to Detection Outlier on one-dimentional datasets.
def find_anomalies(data):
    #define a list to accumlate anomalies
    anomalies = []

    # Set upper and lower limit to 3 standard deviation
    random_data_std = std(random_data)
    random_data_mean = mean(random_data)
    anomaly_cut_off = random_data_std * 3

    lower_limit  = random_data_mean - anomaly_cut_off
    upper_limit = random_data_mean + anomaly_cut_off
    print(lower_limit)
    # Generate outliers
    for outlier in random_data:
        if outlier > upper_limit or outlier < lower_limit:
            anomalies.append(outlier)
    return anomalies
```

Boxplots:

The quantiles of numerical data are represented graphically in box graphs. It's a basic yet efficient method of displaying outliers. Consider the bottom and upper whiskers to be the data distribution's bounds. Outliers or aberrant data points are those that appear above or below the whiskers. The following is the code for creating a box plot:

source code

```python
import seaborn as sns
import matplotlib.pyplot as plt

sns.boxplot(data=random_data)
```

Boxplot Anatomy:

The boxplot graphs are constructed using the Interquartile Range (IQR) idea. IQR is a statistical concept that divides a dataset into quartiles to quantify statistical dispersion and data variability.

To put it another way, each dataset or set of observations is separated into four predetermined intervals depending on the data values and how they relate to the overall dataset. The data is divided into three points and four intervals by a quartile.

DB Scan Clustering:

DB Scan is a clustering method that is used to group data. With single or multi-dimensional data, it may also be employed as a density-based anomaly detection tool. Outliers can also be detected using other clustering techniques such as k-means and hierarchical clustering. I'll show you an example of how to use DBScan in this case, but first, let's go over some key ideas. DBScan has important concepts:

Core Points: We need to look at some of the hyper parameters used to create the DBScan operation to grasp the key points. min samples are the first hyper parameter (HP). Simply said, this is the bare minimum of core points required to construct a cluster. Eps is the second most significant HP. The greatest distance between two samples required for them to be regarded in the same cluster is eps.

Border Points: are in the same cluster as core points, although they're a long way from the cluster's centre.

source code

```python
from sklearn.cluster import DBSCAN
seed(1)
random_data = np.random.randn(50000,2)  * 20 + 20

outlier_detection = DBSCAN(min_samples = 2, eps = 3)
clusters = outlier_detection.fit_predict(random_data)
list(clusters).count(-1)
```

Isolation Forest:

Isolation Forest is a family of unsupervised learning algorithms that includes ensemble decision trees. This strategy is distinct from all prior approaches. All of the preceding ones attempted to locate the data's usual zone, then labelled anything outside that specified zone as an outlier or anomaly.

This procedure is a little different. Instead of profiling and building normal points and areas by assigning a score to each data point, it directly isolates abnormalities. It takes use of the fact that anomalies are rare data items with attribute values that deviate significantly from those of regular cases. This approach works well with datasets with many dimensions, and it's shown to be a good approach to spot abnormalities.

source code

```
from sklearn.ensemble import IsolationForest
import numpy as np
np.random.seed(1)
random_data = np.random.randn(50000,2)  * 20 + 20

clf = IsolationForest( behaviour = 'new', max_samples=100, random_state = 1, contamination= 'auto')
preds = clf.fit_predict(random_data)
preds
```

Robust Random Cut Forest:

Amazon's unsupervised technique for finding abnormalities is the Random Cut Forest (RCF) method. It also operates by assigning an anomaly score. The data point is considered “normal” if the score value is low. High values indicate the presence of a data anomaly. The criteria of "low" and "high" vary depending on the application, but in general, scores greater than three standard deviations from the mean are deemed abnormal.

What does performance testing measure?

Performance testing may be used to assess various success elements, including reaction times and possible mistakes. You may confidently detect bottlenecks, defects, and blunders with these performance statistics in hand and determine how to optimise your application to eliminate the problem (s). Speed, response times, load times, and scalability are the most prevalent performance testing difficulties.

Excessive Load Times

The amount of time necessary to start an application is known as excessive load time. To provide the greatest possible user experience, any delay should be as brief as possible — a few seconds at most.

Poor Response Times

The amount of time between a user putting information into an application and receiving a response to that action is known as poor response time. Users' interest in the program is considerably reduced by long response times.

Limited Scalability

Limited scalability refers to a difficulty with an application's capacity to adapt to varied amounts of users. For example, the program works well with a small number of concurrent users but degrades as the number of users grows.

Bottlenecks

Bottlenecks are systemic bottlenecks that reduce an application's overall performance. Hardware issues or bad coding are the most common causes.

What is the process for performance testing?

The methodology used in performance testing might vary greatly, but the goal of the test remains the same. It can assist you in demonstrating that your software system fulfils pre-defined performance standards. It may also be used to compare the performance of two different software systems. It can also assist you in identifying areas of your software system that are causing it to perform poorly.

Identify your testing environment

Understand your physical test environment, production environment, and available testing technologies. Before you begin the testing process, learn about the hardware, software, and network settings that will be used. It will assist testers in developing more efficient tests. It will also aid in the identification of potential issues that testers may face throughout performance testing methods.

Identify the performance acceptance criteria

This covers throughput, response times, and resource allocation goals and limits. Outside of these aims and limits, it's also vital to develop project success criteria. Because project requirements typically do not offer a diverse collection of performance benchmarks, testers should be given the authority to create performance criteria and targets. There may be none at all at times. Finding a comparable application to compare to is a useful method to set performance targets when possible.

Plan & design performance tests

Determine how end users' use is likely to differ and create key scenarios to test all probable use cases. A variety of end-users must be simulated, performance test data must be planned, and metrics must be defined.

Configuring the test environment

Before starting the test, set up the testing environment. Arrange tools and other resources as well.

- Implement test design
- Create performance tests following your test plan.
- Run the tests
- Run the tests - Run the tests and keep an eye on them.
- Analyse, tune and retest

Compile, evaluate, and disseminate test results. Then fine-tune and test again to determine whether performance has improved or decreased. Stop when the CPU is the bottleneck, as improvements tend to be less with each retest. Then you might want to think about upgrading CPU power.

Example Performance Test Cases

- When 1000 people access the website simultaneously,make sure the response time is less than 4 seconds.
- When network connectivity is sluggish, check that the Application Under Load's response time is within an acceptable range.
- Before the application breaks, check the maximum number of users it can manage.
- When 500 records are read/written at the same time, check the database execution time.
- Check the application's and database server's CPU and memory utilisation under high loads.
- Check the application's reaction time under low, medium, moderate, and heavy load circumstances.

Throughput in Performance Testing

One of the most important measures in performance testing is throughput. It's used to determine how many requests a program can handle per second, minute, or hour.

Every test strategy, on the whole, has a throughput objective. The more realistic it is, the more accurate the result will be.

So, what are the most significant aspects in determining appropriate load characteristics?

- The number of users and their profiles. Whowill be able to use your software or service: customers, purchasers, or administrators?
- Scenarios of behaviour Various users may d odifferent activities, such as purchasing things, filling out forms, or checking the status of a service.
- There are pauses and delays. We all need sometime to process what we've learned. Despite these pauses, the system continues to scan sessions and shut ports.
- Types of connections Different types of network connections impact how the system responds and how users interact with the program.

Data Processing

What Is Data Processing?

Any company cannot use data in its basic form. The method of gathering raw data and converting it into useable information is known as data processing. A team of data scientists and data engineers at a company normally does it in a step-by-step approach. The raw data is gathered, filtered, sorted, processed, Analysed, and stored before being displayed in a usable way.

Data processing is critical for businesses to develop better business strategies and gain a competitive advantage. Employees can comprehend and use the data by translating it to an understandable format such as graphs, charts, and texts.

Data Processing Cycle

The data processing cycle is made up of several phases in which raw data (input) is fed into a process (CPU) that generates actionable insights (output). Each step is performed in a specified order, although the procedure is repeated cyclically. The output of the first data processing cycle can be saved and used as the input for the subsequent cycle.

Collection

The initial stage in the data processing cycle is to acquire raw data. The type of raw data gathered has a significant influence on the final product. As a result, raw data should be collected from defined and accurate sources for the conclusions to be legitimate and useable. Money numbers, website cookies, a company's profit/loss accounts, user activity, and so on are examples of raw data.

Preparation

The act of sorting and filtering raw data to remove unneeded and erroneous data is known as data preparation or data cleaning. The raw data is reviewed for mistakes, duplicates, miscalculations, and missing data before being translated into a format that may be used for further analysis and processing. This is done to guarantee that the processing unit only receives the highest-quality data.

Input

The raw data is transformed to a machine-readable format and sent into the processing unit in this stage. This can take data entry via a keyboard, scanner, or any other type of input device.

Data Processing

The raw data is exposed to numerous data processing technologies to obtain a desirable outcome, including machine learning and artificial intelligence algorithms. Depending on the source of data being processed (data lakes, online databases, linked devices, etc.) and the intended use of the output, this phase may vary slightly from process to process.

Output

Finally, the data is sent to the user in an understandable format, such as graphs, tables, vector files, audio, video, and papers. In the following data processing cycle, this output can be saved and subsequently processed.

Storage

Storage is the final phase in the data processing cycle when data and metadata are saved for later use. This enables easy access to and retrieval of information whenever needed and immediate use of the information as input in the next data processing cycle.

Data Processing Methods

Manual, mechanical, and electronic data processing are the three basic types of data processing.

Manual Data Processing

Data is manually handled in this data processing approach. Without using any other technological equipment or automation software, the whole process of data gathering, filtering, sorting, calculating, and other logical activities are carried out entirely by humans. It is a low-cost approach that requires little to no instruments, yet it results in many mistakes, significant labour expenses, and a lot of time.

Mechanical Data Processing

Data is mechanically processed with the use of gadgets and machinery. Simple gadgets such as calculators, typewriters, and printing presses are examples. This technique may be used to do simple data processing activities. It has fewer mistakes than human data processing, but this approach has become more sophisticated and demanding as the amount of data has grown.

Electronic Data Processing

Modern technology, such as data processing software and program, are used to process data. The program is given a set of instructions to process the input and produce results. This approach is the most costly, but it offers the quickest processing rates as well as the highest level of output dependability and precision.

Sub-Component Performance

It determines how the data's various components are doing. It's important to understand how a query or a map-reduce operation works. Consider how quickly messages are indexed and digested. A code sample demonstrating a simple query of 10 revenue-generating goods is shown below:

```
Most popular product categories
select c.category_name, count(order_item_quantity) as count
from order_items oi
inner join product p on oi.order_item_product_id = p.product_id
inner join categories c on c.category_id = p.product_category_id
group by c.category_name
order by count desc
limit 10;
```

As a consequence, data testing methodologies are persuasive, and data testing outcomes are valuable. They are used by businesses to earn income. The term "big data" is no longer a buzzword, as it is quickly becoming a need. According to Garter, 70% of all businesses will rely on big data for business intelligence in the next several years.

What will I learn?

In this chapter, you will learn about the following:

1. Organise obtained big data sets in a retrievable format
2. Confirm that big data is accurate, up-to-date, and comprehensive
3. Securely store big data and data capture report according to organisational procedures, legislative requirements, and industry practices

Test Case Design

The way you put up your test cases is referred to as test case design. It's critical that your tests are well-designed, or else you risk missing flaws and problems in your product during testing.

To test the functioning and various aspects of your program, you may utilise various test case design methodologies. Good test cases guarantee that every component of your program is thoroughly examined so that any flaws may be identified and corrected.

What are the types of test case design techniques?

The major goal of test case design approaches is to use effective test cases to test the software's functions and features. The three primary types of test case design methodologies are as follows.

Features and formats of common big data sources,

1. Specification-Based techniques
2. Structure-Based techniques
3. Experience-Based techniques

Specification-Based or Black-Box techniques

This methodology uses the external description of the program to develop test cases, such as technical specifications, design, and customer needs. Testers can use this methodology to create test cases that cover all aspects of a project. There are five sorts of specification-based or black-box test case design methodologies. The following are the categories:

Boundary Value Analysis (BVA)

This method is used to investigate faults at the input domain's edge. BVA detects any input problems that might cause the software to stop working properly.

Equivalence Partitioning (EP)

The test input data is partitioned into several classes with a comparable number of data in Equivalence Partitioning. After that, test cases are created for each class or division. As a result, the number of test cases is reduced.

Decision table testing

Test cases are created using this methodology based on decision tables that are created using various combinations of inputs and outputs based on various circumstances and situations according to various business rules.

State transition diagrams

The program under test is viewed as a system with a finite number of states of various sorts in this methodology. A system of rules governs the transition from one state to the next. The rules specify how the system reacts to various inputs. This strategy can be used on systems that have specific processes.

Use case testing

A use case is a description of how a user uses the product in a certain way. The test cases in this method are meant to run various business scenarios and end-user functionality. Use case testing aids in the identification of system-wide test cases.

Structure-Based or White-Box techniques

The structure-based or white-box methodology creates test cases based on the software's underlying structure. This method thoroughly checks the code that has been produced. Developers with a thorough understanding of the software code, its internal structure, and design assistance in creating test cases. There are five different types of this approach.

Statement testing & coverage

This method entails running all of the source code's executable statements at least once. According to the specified requirement, the proportion of executable statements is determined. For checking test coverage, this is the least desired measure.

Decision testing coverage

This approach, also known as branch coverage, is a testing approach that ensures all reachable code is run by executing each of the potential branches from each decision point at least once. This aids in validating all of the code's branches. This ensures that no branch causes the programme to behave in an unanticipated way.

Condition testing

Each Boolean expression is predicted as TRUE or FALSE in condition testing, also known as predicate coverage testing. All of the test results are tested at least once. This method of testing ensures that the code is completely covered. The test cases are written in such a way that the condition results are simple to implement.

Multiple Condition Testing

The goal of multiple condition testing is to test every possible combination of situations to achieve 100% coverage. Two or more test scripts are necessary to assure comprehensive coverage, which necessitates more effort.

All Path Testing

The source code of a program is used to discover every executable path in this method—this aids in the identification of all flaws in a given code.

Experience-Based techniques

To grasp the most critical aspects of the program, these methodologies rely heavily on the tester's experience. The abilities, knowledge, and competence of the persons engaged determine the outcomes of these procedures. The following are examples of experience-based techniques:

Error Guessing

The testers use this methodology to predict faults based on their previous experience, data availability, and product failure knowledge. The testers' abilities, intuition, and experience all play a role in error guessing.

Exploratory Testing

Without any formal documentation, this methodology is used to test the programme. A minimum amount of time allowed for testing and a maximum amount of time allowed for test execution. The test design and execution are done simultaneously in exploratory testing.

Formulating a Big Data

Organisations should be able to implement a well-defined enterprise Big Data strategy. To do this, businesses may use the 5-step process shown below to develop their Big Data strategy:

1. Define business objectives
2. Execute a current state assessment
3. Identify and prioritise Use Cases
4. Formulate a Big Data Roadmap
5. Embed through Change Management

Define business objectives:

To properly exploit Big Data in any company, it is first required to completely comprehend the enterprise's corporate business objectives. What factors contribute to a company's success? Meeting or exceeding corporate Key Performance Indicators generally results in increased revenues and profitability (KPIs). Begin by learning how a company succeeds before looking at how Big Data technology and solutions might improve future performance.

Because the fundamental goal of Big Data is to extract value by exploiting data, it should be aligned with corporate business objectives and address major business concerns.

The necessity of including important corporate stakeholders in the identification of corporate objectives cannot be overstated. Ensure that these stakeholders are included from the beginning and offer valuable feedback regularly. The following are key stakeholders to consider in this initial step:

- Sponsors from the executive suite. It's impossible to overestimate the value of obtaining and partnering with executive sponsors.

Their assistance is critical during the ups and downs of developing and implementing the Data Strategy.

- On the team, there is the right talent. To determine the correct company objectives, individuals with the proper ability and skill sets must be involved. Both internal and external consultants should be considered.
- Troublemakers in the making. There will be certain "stakeholders" in any project or endeavour who are either consciously or unwittingly hostile to change. Knowing who they are and what drives them will aid you later in the process.

Execute a current state assessment:

The major goal of this stage is to evaluate the enterprise's present business processes, data sources, data assets, technological assets, capabilities, and policies. This activity aims to aid in gap analysis between the current state and the planned future state.

If the goal of the data strategy is to gain a 360-degree perspective of customers and future customers, the current state evaluation will cover any business process, data assets, architecture, capabilities (including business and IT), and departmental rules that affect consumers. Typically, a current state evaluation entails a series of interviews with staff involved in client acquisition.

Identify and prioritise use cases:

Develop Use Cases that connect with the business objectives from step 1 to envisage how predictive analytics, prescriptive analytics, and ultimately cognitive analytics may help the company accelerate, optimise, and continually learn. To understand how Big Data may help you achieve your business goal, write down each Use Cases.

Use Cases are a simple and effective approach to explain how Big Data technology and solutions may help companies achieve their objectives. Following the development of the Use Cases, the next stage is to prioritise them all based on their business effect, budget, and resource needs. Enterprises may use this exercise to determine which Big Data efforts bring the most business value.

A Prioritisation Matrix is one of the most effective techniques to prioritise Use Cases. The Prioritisation Matrix facilitates discussion and debate between business and IT stakeholders in identifying the "right" Use Cases to launch a Big Data initiative, i.e., those Use Cases with meaningful business value (from the perspective of business stakeholders) and reasonable feasibility of successful implementation.

Formulate a Big Data Roadmap:

The following stage is likely to be the most time-consuming and controversial, and it will undoubtedly consume the bulk of the time spent developing a data strategy. The Roadmap may be created based on the existing capability status assessment (step 2) and the recognised and prioritised Big Data Use Cases (step 3). The Big Data Roadmap lays out which initiatives (or Use Cases) will be completed first, as well as which capabilities (knowledge, tools, and data) will be expanded over the following three years.

The Roadmap should focus on identifying gaps in data architecture, technology and tools, processes, and, of course, people, with the intended future state in mind (skills, training , etc.).

Sponsors and stakeholders will play an important role in prioritising these projects. This phase concludes with a strategy for implementing the priority Big Data activities.

Embed through Change Management:

Although Change Management (engaging people's hearts and minds) is not formally part of the Big Data Strategy formulation, it will significantly influence the success or failure of a Big Data strategy.

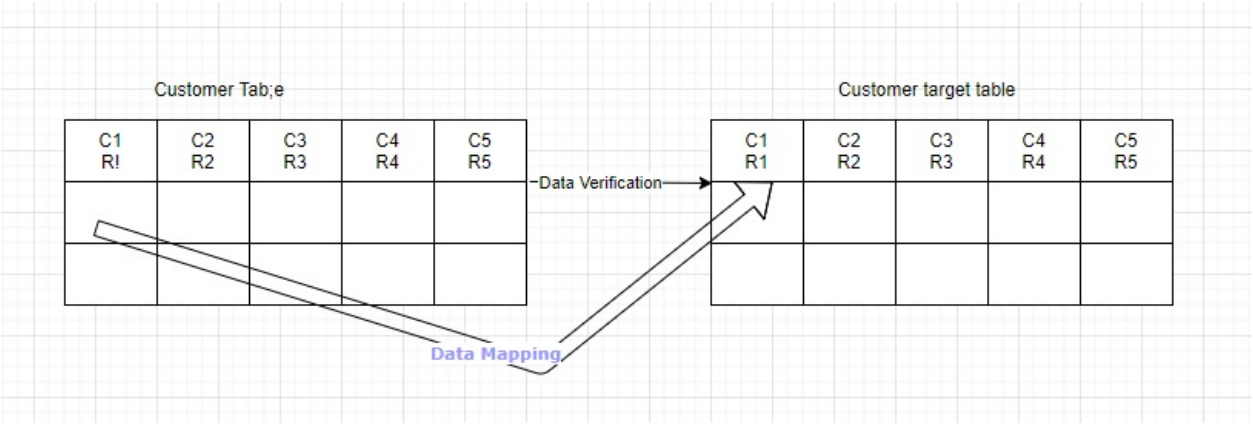
Organisational change, cultural change, technical change, and business processes should all be included in change management. Data governance, or the total management of data availability, accessibility, integrity, and security, has become an important part of change management. Any change management program should include appropriate incentives and continuous measurements.

Test Scenarios And Test Cases

We need to produce Test Scenarios and Test Cases when we finish the Test Planning, especially for Big Data Testing, which requires a few papers and the requirement document. What more do we require in addition to this requirement document?

We'll need the Requirement Document, which provides the Client's requirements and the Input Document, which provides the Data Models. This data will be available in the Data Models, including the DataBase Schemas, tables, and relationships.

Also, we have Mapping Documents. For example, in Relational DataBases, we have certain Tables, and after loading the data through ETL in the Data Warehouse in HDFS, what mapping do we need to make? i.e. Mapping Data Type Mapping Data Type Mapping Data Type Mapping Data Type Mapping Data Type Mapping



We have certain columns in this Customer Table, and we have certain columns in the CUSTOMER TARGET Table, as indicated in the diagram. The data was spilled from the Customer table to the CUSTOMER TARGET table, i.e. from source to target.

Then we need to double-check the mapping, such as the data in the Source Table, which is the Customer Table's Column 1 and Row 1 and is designated as C1R1, and the same data should be mapped in C1R1 of the CUSTOMER TARGET Table. This is referred to as mapping.

What will we do if we don't know what all the Mappings need to be verified? As a result, the Mapping Document will include these Mappings. The Customer will provide several mappings in the Mapping Document.

We've also requested a Design Document, which is necessary for both the Development and QA Teams since the Design Document will detail what types of Map Reduce Jobs the customer will implement and what types of MapReduce Jobs receive inputs what types of **MapReduce Jobs produce outputs.**

Similarly, if we're using HIVE or PIG, what are all of the UDFs the customer has generated, as well as all of the inputs and outputs they'll produce, and so on.

To prepare Test Scenarios and Test Cases, we need to have all these Documents by hand:

- Requirement Document
- Data Model
- Mapping Document
- Design Document

These can differ from one organisation to the next, and there is no need to have all of these papers. Depending on the intricacy of the project, corporate schedules, and other factors, we may have all papers, just two or three papers, or we may need to rely on only one document.

Tools for testing

Big Data integrates with Hadoop, Teradata, MongoDB, AWS, and other NoSQL technologies using various automated testing methods. To support continuous delivery, it must be integrated with dev operations. These tools must have a robust reporting capability and be scalable, adaptable to frequent changes, cost-effective, and dependable. These technologies are mostly used in big data testing to automate repetitive activities.

- HDFS (Hadoop Distributed File System)
- Hive
- HBase
- MapReduce
- HiveQL
- Pig Latin

Process	Tools Description
Data Ingestion	Zookeeper, Kafka, Sqoop
Data Processing	MapR, Hive, Pig
Data Storage	Amazon S3, HDFS
Data Migration	Talend, Kettle, CloverDX

Challenges faced in Testing Big Data

- Big Data Testing is a difficult procedure that needs the assistance of a highly qualified official.
- The processes for automated Big Data Testing are predetermined and unsuitable for unanticipated mistakes.
- Latency in testing is caused by virtual machine latency, and controlling multimedia is a hassle.
- One of the most difficult aspects of testing is dealing with large amounts of data.
- For many platforms, a test environment and automation should be created.
- Because each component is from a distinct technology, it must b etested separately.
- End-to-end testing is impossible to do with a single tool.
- Designing test scenarios necessitates a high level of scripting.
- To improve performance and test important areas, customised solutions are necessary.

Quality assuring output

Plan Quality Management and Control Quality are wedged between Quality Assurance and Plan Quality Management. The flow is different in reality. The PMP Certification Exam will test your knowledge of project quality assurance directly.

Perform Quality Assurance: Outputs

Quality audits and process improvement efforts aim to lower quality costs and/or improve customer satisfaction. The Integrated Change Control process will be used to implement any corrective actions found during the quality audit and any chances to enhance procedures. Change requests can be made in a variety of ways.

- Changing a policy
- Changing a procedure
- Changing a process
- Corrective action
- Preventive action
- Changing the quality management plan

You may need to alter the project management strategy, project papers, and organisational process assets based on the outcomes of the Perform Quality Assurance procedure.

Test script for big data

What is a Test Script?

Test scripts are a line-by-line description of the system transactions that must be done to validate the application or system under test. Each step should be listed in the test script, along with the intended outcomes.

This automation script enables software testers to thoroughly test each stage on a variety of devices. The actual items to be executed and the expected results must be included in the test script.

How to Write a Test Script

There are three different ways to create a test script:

1. Record/playback:

Instead of simply recording the user's activities, the tester must write any code in this function. However, the tester may be required to code in order to correct errors or fine-tune the automation behaviour.

Because you already have the whole code, this method is easier than building a comprehensive test script from scratch. It's most commonly found in a simplified programming language like VBScript.

2. Keyword/data-driven scripting:

There is a clear distinction between testers and developers with this strategy. The tester defines the test using keywords rather than the underlying code in data-driven scripting.

The developers' task here is to implement the test script code for the keywords and update it. As a result, the tester does not have to be concerned about the system when using this method. For any new functionality, you want to test automatically. However, they will heavily rely on development resources.

3. Writing Code Using the Programming Language:

If you prefer to construct test scripts this way, you will usually record or playback the results and construct a simple script.

However, as a tester, you will eventually need to learn how to create basic scripts and record/playback. Even though your application is created in Java, you have the option of choosing your programming language.

However, it does not imply that you must create your test scripts in Java, which is a challenging language to master. Instead, use a simpler language like JavaScript or Ruby to construct your test scripts (or any easier language you wish to use).

Example of a Test script

For instance, to test a website's login feature, your test script may execute the following:

- Specify where the "Username" and "Password" fields on the login screen should be found by the automation tool. Let's suppose we are going to go by their CSS element IDs.
- Go to the homepage of the website and click the "login" option. Check that the Login screen is visible and the "Username" and "Password" columns.
- Next, input the login "Robert" and password "123456", then locate and click the "Confirm" button.
- They must describe how a user may get the title of the Welcome screen that shows after logging in, for example, by its CSS element ID.
- Make sure the Welcome screen's title is displayed.
- Take a look at the title of the web screen
- "Welcome, Robert", write in the title text.
- If the headline wording matches the expectations, the test was successful. Otherwise, an album that fails the test.

Isolating sub-standard data and correcting data acquisition paths

Different data processing architectures for big data have been proposed to address the different characteristics of big data. Data acquisition has been understood as the process of gathering, filtering, and cleaning data before the data is put in a data warehouse or any other storage solution. The acquisition of big data is most commonly governed by four of the Vs: volume, velocity, variety, and value. Most data acquisition scenarios assume high-volume, high-velocity, high-variety, but low-value data, making it important to have adaptable and time-efficient gathering, filtering, and cleaning algorithms that ensure that the data-warehouse analysis actually processes only the high-value fragments of the data.

Key Insights for Big Data Acquisition

The foundation of data acquisition across the various architectures for big data processing boils down to obtaining data from distant information sources to store it in scalable, big data-capable data storage. Three essential components are necessary to attain this goal:

- Protocols that allow for the collection of data from any sort ofdistributed data source (unstructured, semi-structured, structured)
- Frameworks for collecting data from a variety of dispersed sourcesutilising various protocols.
- Technologies that allow the frameworks to save the data they retrievein a persistent format.

Social and Economic Impact of Big Data Acquisition

This development presents various possibilities and difficulties for businesses in developing new business models and optimising current operations, resulting in market advantages. The four Vs-driven tools and methodologies for dealing with large data may be used to better user-specific advertising or market research in general. In the energy industry, for example, smart metering technologies are being tried. In addition, when used in conjunction with new billing systems, these methods might be useful in other industries such as telecommunications and transportation.

Big data has already had an impact on many firms and can do so across all industries. While there may be some technical difficulties, the influence on management, decision-making, and even business culture will be significant.

However, there are still certain limitations. These systems and technologies must, in particular, handle privacy and security considerations. Although many systems create and collect vast volumes of data, only a tiny portion of that data is actively employed in business activities. Furthermore, many of these systems do not meet real-time needs.

Future Requirements and Emerging Trends for Big Data Acquisition

Big data collection software must cope with high-volume, diverse, and real-time data. As a result, data acquisition tooling must assure a high throughput. This implies that data can originate from a variety of sources (social networks, sensors, web mining, logs, and so on) and have various forms, or it can be unstructured (text, video, photos, and media files) and arrive at a rapid rate (tens or hundreds of thousands of events per second). As a result, providing frameworks and tools that provide the appropriate throughput for the situation at hand without losing any data is the primary difficulty in obtaining big data.

The following are some of the developing issues for large data collecting in this context:

- Tools that offer some form of input data to the system, such as socialnetworks and web mining techniques, sensor data gathering software, logsinjected regularly, and so on, are frequently used to start data gathering. Thedata acquisition process usually begins with a single or several endpoints fromwhich the data originates. These endpoints might take several technologicalforms, such as log importers, Storm-based algorithms, or even data collectionAPIs to the outside world for data injection via RESTful services or otherprogrammatic APIs. As a result, any technical solution that attempts to collectdata from many sources should be able to handle this wide variety of data.
- Both the historical and real-time levels provide ways to integratedata collecting with data pre-and post-processing (analysis) and storage. Toaccomplish so, the data capture tools should access the batch and real-timeprocessing tools (e.g., Storm and Hadoop).
- This is done in a variety of ways. Apache Kafka, for example, offers aPublish-subscribe system to which both Hadoop and Storm may subscribe, makingthe messages received available to them. On the other hand, Apache Flume takesa different method, storing data in a NoSQL key-value store and pushing it toone or more receivers to assure velocity.
- The gathering of material (photos, video) is a considerabledifficulty, but the processing and preservation of video and photos are evenmore difficult.
- To appropriately and efficiently blend data from diverse sources whenprocessing, data diversity necessitates processing the

semantics in the data.

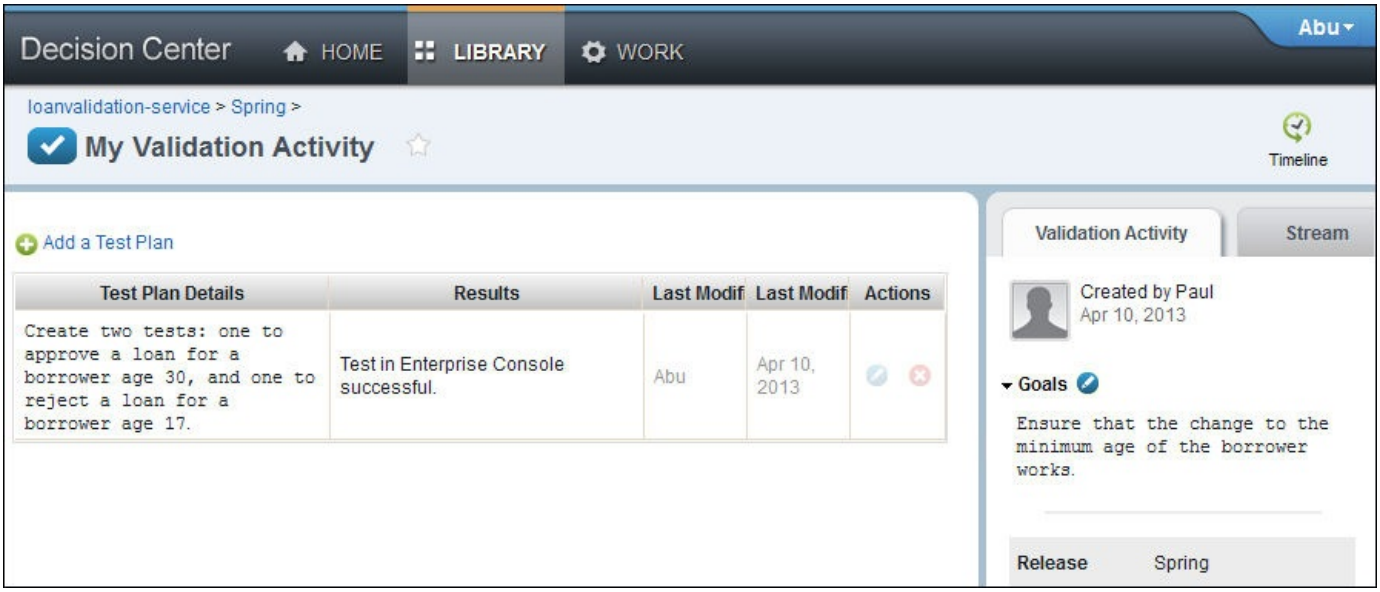
- The present state-of-the-art provides a range of open-source andcommercial tools and frameworks for performing post- and pre-processingobtained data. When developing a proper data acquisition plan, the key aim isto understand the system's requirements regarding data volume, diversity, andvelocity and then to choose the optimal instrument to assure the acquisitionand necessary throughput.

Substandard Data

Data on scrap paper, incorrectly transcribed data, inappropriate error repairs, data that is inconsistent with protocols, and other issues are instances.

Validation activities are generated and stored in the Business console and are used to track and manage a test plan for the release and the results.

The following image shows an example of a validation activity on a spring release:



You define what tests assure a proper validation. For example, you can use the Enterprise Console to run test suites and simulations on the release and incorporate the report into the validation activity results.

When all validation activities are completed, the release can be approved and completed, at which point deployment can occur.

Validation activity governance

As long as the release is In Progress, all users can create a validation activity for the release and assign an owner to the activity.

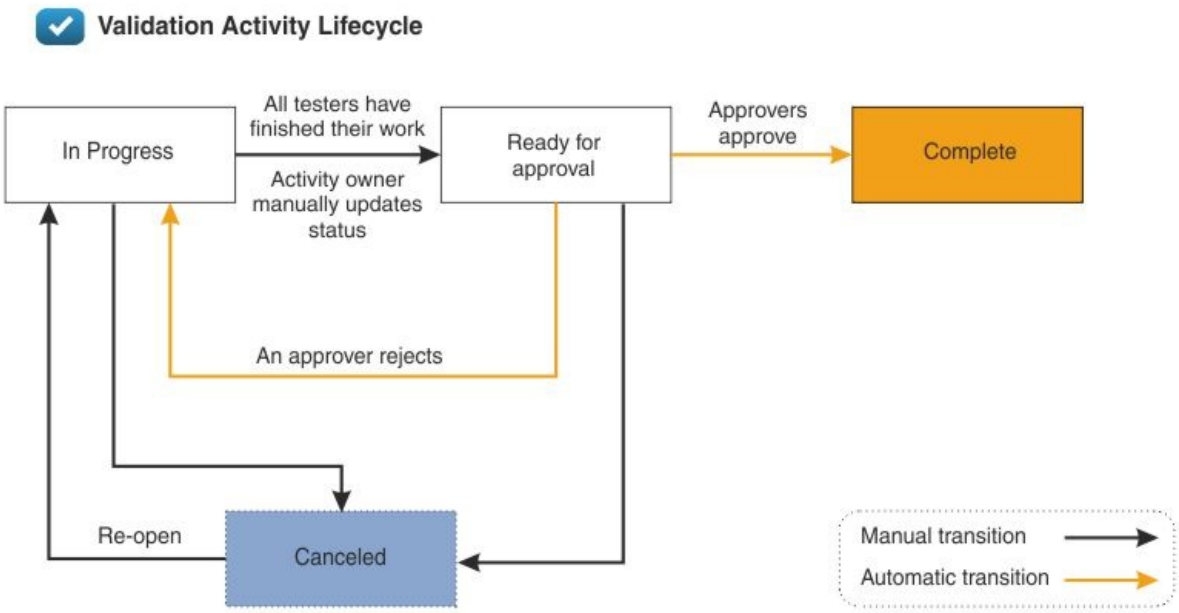
Then, as long as the validation activity is not Complete, the owner of the activity can:

- Change the owner of the activity.
- Change the due date of the activity.
- Change the goals of the activity.
- Assign approvers and testers to the activity

Testers run different tests aimed at validating a release and note the result in the test plan. Once they finish their work and all the change activities of the release are Complete, testers change their status to Finished. When all the testers finish their work, the owner of the validation activity sets its state to Ready for Approval, at which point the approvers approve or reject the activity.

When all the approvers have approved the validation activity, Decision Center sets the state of the validation activity to Complete.

The following diagram shows the lifecycle of a validation activity, with manual transitions being done by the owner of the activity:



The following table shows the most frequent user operations on a validation activity, the role required to do the operation, the precondition to this operation, and the state of the validation activity after:

User operation	Role	The precondition to the operation	State of activity after operation
Create activity	-	-	In Progress
Cancel activity	Owner	-	Cancelled
Reopen activity	Owner	Activity is in a Canceled state	In Progress
Proceed to approval	Owner	Testers have finished working	Ready for Approval
Finish working	Tester	Tester working	In Progress
Resume working	Tester	Tester has finished working	In Progress
Approve changes	Approver	Activity is in Ready for Approval state	Complete
Reject changes	Approver	Activity is in Ready for Approval state	In Progress

You can use a Validation in a pipeline to ensure the pipeline only continues execution once it has validated the attached dataset reference exists, that it meets the specified criteria, or timeout has been reached.

```
{
  "name": "Validation_Activity",
  "type": "Validation",
  "typeProperties": {
    "dataset": {
      "referenceName": "Storage_File",
      "type": "DatasetReference"
    },
    "timeout": "7.00:00:00",
    "sleep": 10,
    "minimumSize": 20
  }
},
{
  "name": "Validation_Activity_Folder",
  "type": "Validation",
  "typeProperties": {
    "dataset": {
      "referenceName": "Storage_Folder",
      "type": "DatasetReference"
    },
    "timeout": "7.00:00:00",
    "sleep": 10,
```


Generate and store results of validation activity and associated supporting evidence according to organisational policies and procedures and legislative requirements.

Generating and storing evidence is an important factor in achieving successful test results. It will be one of, if not the, most important factors in determining whether or not what you are working on will have an impact.

While evidence-based approaches to humanitarian innovation are recognised as important, most people struggle to generate and use evidence strategically, explicitly, and systematically toward developing innovative humanitarian solutions.

Without evidence, you risk developing solutions that are, at best, unsuitable for users, waste time and money to develop, are insignificant in addressing needs, or, at worst, may cause unintentional harm. As a result, you must develop a process for generating and gathering evidence throughout the innovation cycle.

Understand the role of evidence and how it can be applied

To better understand the relationship between evidence and results, consider the following three key concepts:

To begin, evidence is not the same as information. Whereas information is defined as the facts and details learned about something through study or experience, evidence can be used to support or refute a hypothesis or claim.

Second, the role of evidence changes throughout the lifecycle of test results. It is not limited to conducting an assessment at the start of a project and evaluating at the end of a pilot. We can think of evidence application as a resource for the following activities:

- Investigating and explaining problems
- Developing insights to guide the creation of new ideas
- Validating and stress-testing initial concepts, as well as fine-tuningthe functionality of viable concepts
- Showing the impact of real-world solutions
- Recording, reflecting on, and sharing learning with the larger community
- Providing strategic decision-making support and articulation at keystages and pivot points throughout the test results process

Third, using evidence to support test results necessitates a procedure. Using evidence for results necessitates the identification of

relevant information and a process for putting that knowledge into action.

Demonstrate improvement and learning

Finally, any new result must demonstrate that it outperforms existing solutions. To do so, you must understand the problem as well as the efficacy of existing solutions and your solution, and you must be able to compare the two. You must also analyse and document your learning and improvement process.

The Problem: Too often, when people begin to develop solutions for a problem, they only have a surface-level understanding of it. The more you understand the problem and the evidence you have, the more likely you are to find or develop a relevant, impactful solution. The more quantitative and qualitative evidence that can be generated toward this goal to prove that there is a problem, the better. The section on Recognition in the Guide delves into this.

Existing solutions

How effective are the current practices and solutions to the problem? Do you have proof of their effectiveness and impact? This is frequently extremely difficult to locate. Several approaches can be taken to investigate this, and we will look at some of them in the Search, Adapt, and Pilot sections of this guide. The more quantitative and robust qualitative evidence that can be generated to demonstrate the performance and impact of existing solutions, the easier a comparative analysis will be.

Your solution

If you have adopted or developed a solution, your results will need to be put through a rigorous evidence-building process. This will require, at the bare minimum, a strong DME/MEAL (Design, Monitoring and Evaluation/Monitoring, Evaluation, Accountability and Learning) process but should have a further research component to it as well.

Comparison between your solution and other solutions:

For your results to be considered successful, you must evaluate your solution's performance and impact on the problem compared to the performance and impact of other solutions to the problem. This should be done using a comparative analysis. By doing so, you are attempting to demonstrate whether or not your solution is superior to other possible solutions to the problem.

Storing Test results

The storage of test results is entirely dependent on the organisation and the tools used. Many organisations use JIRA with the Zephyr plug-in, HP QC, Silk Central, and other similar tools. In addition, there are free test results management tools on the market.

An ideal system to store test results should have the following characteristics:

- Allows storing requirements
- Allows for the storage of test cases as well as their grouping intosuites and sets
- Allows for the storage of test results on a per-build basis; a failedtest case should be linked to a defect
- Allows for the storage of defect/issue information
- Allows for the storage of build information
- Has code repository (git is preferred)
- Has a Wiki or a similar system for storing knowledge?
- Has integrations with other systems and/or a good API, allowing us tointegrate it ourselves

Here are some tools for managing test results for your convenience. Choose based on the features they provide and whether you have any automation tools that must be paired with them.

- Jira
- Github
- CircleCI

We've found Jira to be more flexible and to have a better UI than Trello, VSTS, and Pivotal Tracker. Jira also works well with the same company's Confluence. Github and gitflow are the lightest code review systems I've seen, and Github is the most popular code storage site.

CircleCI is the most recent Continuous Integration system we've used, and we much prefer it to Jenkins (the market leader) and VSTS (Microsoft offering).

Organisational policies and procedures relating to testing big data sources

How to Implement Successful Big Data in an organisation

Assembling and obtaining raw big data:

Collecting raw data (also known as primary data) is the beginning point for any data analysis. Once the RD (Raw Data) has been acquired, it is processed into Information, which may then be transformed into Knowledge later in the analysis process. This White Paper aims to highlight the basic concepts and problems of RD collection, as well as how these concepts and problems are developing in light of new technology and business processes.

Raw Data Collection:

All of the previous ways to RD collection followed a five-step process:

1. Establishing what information/data needs to be collected
2. Setting a timeframe for the collection
3. Establishing a method for the collection
4. Collecting the Data
5. Sorting the Data

For many years, Steps 1–4 were thought to be simple to complete. There were standard collecting techniques for each of the data kinds that could be implemented without much trouble. Many sectors had established industry-wide data gathering and presentation standards. Banking and finance, travel, resource sectors, and international shipping and trade are some of the businesses where data collecting has been standardised to a large extent. Data scientists may create so-called "templates," which they could then use as needed. In the rare case where no template appeared to be suitable for data gathering due to project-specific constraints, new templates could generally be created by re-engineering existing templates. It was seldom necessary to create templates from scratch.

Assembling

Cherish your data

Teal advises, "Keep your raw data raw: don't change it without a duplicate." She suggests keeping your data in a secure location with automatic backups and access by other laboratory members while adhering to your institution's permission and data protection policies.

Because you won't need to view these files frequently, Teal advises, "you may choose storage choices where accessing the data may be more expensive, but storage costs are minimal" — such as Amazon's Glacier service. You may even keep multiple hard discs in separate places to save the raw data. Large data file storage expenses can add up quickly, so plan.

Visualize the information

The practice of expressing data visually and understandably so that a user may better comprehend it is known as information visualisation. Information visualisation examples include dashboards and scatter plots. Information visualisation helps users to efficiently and effectively derive insights from abstract data by providing an overview and identifying pertinent relationships.

Show your workflow

This entails documenting the whole data pipeline, including the data version you used, the data clean-up and quality-checking stages, and any processing code you used. This type of information is crucial for recording and repeating your procedures. To record what he enters into the command line, Eric Lyons, a computational biologist at the University of Arizona in Tucson, uses the video-capture programme asciinema, although lower-tech methods can also work. He recalls many of his coworkers taking images of their computer screens and posting them to the lab's Slack channel, and instant-messaging network.

Use version control

a peer-to-peer network that allows you to share and version files of any size for free. The system keeps track of all the operations you conduct on your file in a tamper-proof log.

The system keeps track of all the operations you conduct on your file in a tamper-proof log.

Record metadata

Metadata must describe how observations were gathered, structured, and structured. Consider which metadata to capture before you begin collecting, and save that information with the data — either in the software tool used to gather the observations or in a README or other dedicated file, according to Lyons.

Capture your environment

You'll also need the same operating system as the tool and all of the same software libraries. As a result, he suggests working in a self-contained computer environment, such as a Docker container, which can be constructed almost anywhere. To record and share their virtual environments, the internet platform Code Ocean (which is based on Docker).

Organisational Policies

Employee and employer duties are outlined in company policies and procedures, which govern the standards of behaviour inside a business. Company rules and procedures are in place to protect both workers' rights and employers' commercial interests. Various policies and procedures define standards about employee behaviour, attendance, dress code, privacy, and other areas connected to the terms and circumstances of employment, depending on the needs of the business.

Employee Conduct Policies

As a condition of employment, an employee conduct policy defines the obligations and obligations that each employee must follow. As a guideline for appropriate employee behaviour, conduct regulations describe correct dress code, workplace safety protocols, harassment regulations, and computer and Internet usage regulations. Employers can use such policies to specify the measures they can use to penalise employees who engage in inappropriate behaviour, such as issuing warnings or terminating them.

Companies are increasingly seeing bullying as a major issue and are beginning to implement measures to address it. The focus of anti-bullying regulations is on persistent aggressive behaviour.

Equal Opportunity Policies

Equal opportunity laws are policies that ensure that employees are treated fairly in the workplace. To encourage equitable behaviour in the workplace, most businesses employ equal opportunity rules, such as anti-discrimination and affirmative action laws. These rules prohibit workers, managers, and independent contractors from behaving inappropriately against others in the business based on their race, gender, sexual orientation, or religious and cultural views.

Attendance and Time Off Policies

Many companies have substance abuse policies that prohibit drug use, alcohol and tobacco products during work hours, on company property or during company functions. These policies often outline smoking procedures employees must follow if allowed to smoke on business premises. Substance abuse policies also discuss the testing procedures for suspected drug and alcohol abuse.

Substance Abuse Policies

Employee adherence to work schedules is governed by attendance policies, which provide standards and criteria. Attendance regulations govern employees' ability to arrange time off or alert superiors of an absence or late arrival. This guideline also outlines the ramifications of failing to stick to a schedule. Employers, for example, may only allow a particular amount of absences within a given time window. According to the attendance policy, employees who skip more days than the firm permits will risk disciplinary action.

Workplace Security Policies

Security policies exist to safeguard the people in an organisation and the physical and intellectual property. Entrance to a facility, such as the usage of ID cards and the processes for signing in a visitor, may be covered by policies. It may be necessary to sign off equipment such as a work laptop or smartphone.

These days, computer security is a top responsibility for businesses. Policies address a wide range of subjects, including how often passwords should be changed, how to report phishing attempts, and how to log in. Personal devices, such as a USB drive brought from home may also be prohibited to avoid the unintentional transmission of computer viruses and malware.

Legislative requirements

GUIDE TO BIG DATA AND THE AUSTRALIAN PRIVACY PRINCIPLES

The OAIC published a draught Guide in May 2016 to assist entities in conducting big data operations in compliance with privacy legislation.

IN BRIEF

- The Office of the Australian InformationCommissioner (OAIC) has produced a draught ‘Guide to Big Data and theAustralian Privacy Principles (Guide) for consultation.
- The OAIC recognises the significance andimportance of big data activities and strives to strike a balance between thebenefits of big data activities and the protection of personal information andprivacy.
- When personal information is acquired,handled, and preserved for the sake of big data activities, the Guide suggestsa number of

steps that entities should take.

The guide

The OAIC published a draught Guide in May 2016 to assist entities in conducting big data operations in compliance with privacy legislation. The draught Guide is aimed at entities that are subject to the Australian Privacy Principles (APPs) set out in the Privacy Act 1988(Cth) (Privacy Act), such as the Federal government and many private sector companies (including those with annual revenues of more than $3,000,000 and health service providers). The draught Guide can also be used as a model for organisations that are not covered by the APPs.

The Guide will not be legally binding until it is finalised, but it will serve as a guide for the OAIC as it fulfils its responsibilities under the Privacy Act.

General recommendations

Prior to engaging in big data operations, the Guide presents two fundamental general pieces of advice to entities.

To begin with, the OAIC suggests that organisations include and embed privacy into their culture, procedures, and systems from the start ('privacy by design). This guarantees that privacy is integrated into an organisation or project rather than being a last-minute consideration. Entities involved in big data projects should incorporate privacy considerations into their plans, including undertaking a privacy impact assessment to identify hazards and make necessary recommendations.

Second, if possible, data gathered and used for big data operations should be de-identified. De-identification removes the information outside the reach of the Privacy Act, allowing a firm to more freely use and maximise the value of the data. The appropriate method of de-identification for the nature of the data, the appropriate uses and disclosures of the de-identified information, the stage at which de-identification should occur, and the cost, difficulty, practicability, and likelihood of the information being re-identified are all relevant considerations for entities.

The Australian privacy principles and key considerations for storing data

The OAIC's application of the APPs to big data shows that businesses engaging in big data activities can do so while adhering to their APP duties, including notifying individuals involved in the big data activities. The OAIC recommends that a company that engages in big data operations do the following:

1. limit the gathering of personal information to the bare minimum required to carry out its big data activities

2. When personal information is used or disclosed for big data activities, it should be clearly stated in a privacy notice, as this is typically a "secondary purpose" to the original purpose for which the data was obtained, as well as whether the data would be used for direct marketing reasons.

3. Individuals should be able to choose which uses, gathers, or disclosures of personal information they consent to and which they do not.

4. Individuals should be notified if their information will be shared with third parties, especially international receivers (which is often the case in big data activities).

5. Individuals should be informed of the specifics of any information obtained from third parties, as well as the fact that such disclosures may occur, as stated in the privacy notices of these third parties.

6. Rather than depending on a single static privacy policy, give more dynamic, multi-layered, and user-centric privacy notices such as "just-in-time" notices, video notifications, and privacy dashboards.

7. Ensure that people may easily opt-out of receiving future marketing communications or request that their information not be used for such purposes. A company should also assess whether its big data activities are aiding direct marketing by other companies to whom it provides data. This is significant since big data activities are frequently carried out to inform direct marketing campaigns.

8. Keep track of the different categories of data being gathered to avoid the danger of utilising or revealing sensitive data for direct marketing purposes without the consent of the individuals.

9. If information needs to be maintained for future big data operations, de-identify it before transmitting it overseas for big data operations (as may occur through the use of overseas cloud or internet-based platforms).

10. Take reasonable precautions to ensure that foreign recipients of information do not violate the APPs since the Australian organisation will be held liable for any violations.

11. Take stringent steps to preserve the quality of data utilised for big data activities, as the data has the potential to become obsolete or erroneous due to the vast volume of data acquired from a variety of third-party sources and often maintained for long periods of time.

12. Implement methods to examine the quality of information, such as keeping track of when it was obtained and verify that third-party sources of information have followed the same standards, procedures, and systems.

13. Take necessary precautions to monitor and safeguard against the security threats posed by big data activities. This is due to the fact that such activities are "honey pots" of valuable and sensitive personal information that are often kept for extended periods of time.

Key takeaways

Because of the nature of big data activities – that is, they involve large sets of data (and may result in the creation of personal information through the aggregation of different data sets) that are frequently sourced from or shared with third parties, sourced originally for different purposes, and retained for long periods of time – they pose potentially greater risks to personal privacy.

Businesses that engage in big data activities (including entities that perform big data analytics or rely on the findings of such analytics) should be aware of the increased risk of violating personal privacy and consult this Guide to ensure that their operations are compliant with the Priva Act.

The draught Guide is consistent in many ways with the OAIC's other guidance publications, such as the APP Guidelines. As a result, many of the draught Guide's suggestions should already be followed by companies engaged in big data activities. The Guide does, however, emphasise key areas that are likely to cause those institutions to reconsider their actions, notably in relation to:

1. In privacy notices, offering more specific information on big data activities without being unduly explicit.

2. When personal data is used for numerous reasons, the degree of choice given to individuals regarding which purposes they accept and which they do not accept.

3. Rather than depending on a single static privacy policy, provide more dynamic, multi-layered, and user-centric privacy disclosures.

4. Enhancing data quality measures, such as keeping track of when personal data was gathered if it was an opinion, and whether it was gathered through creation.

5. When third parties are involved, and evaluation of their privacy policies, notices, practises, processes, and systems are required.

The government has asked for public feedback on the draught Guide, which must be submitted by July 25, 2016.

Legal notice

What will I learn?

In this chapter, you will learn about the following:

1. Perform data cleansing on big data sample following testing according to industry practices and organisational procedures
2. Collate validated output of testing, confirming absence of big data corruption in sample
3. Recommend configuration optimisation changes based on performance testing results
4. Communicate final sample results to required personnel

Performing data cleansing following extract, transform and load (ETL) testing

ETL is a process that extracts data from several source systems, transforms it (by applying computations, concatenations, and other operations), and then inserts it into a Data Warehouse system. ETL stands for Extract, Transform, and Load.

It's easy to believe that building a data warehouse is simply pulling data from many sources and loading it into a database. This is far from the case, and a sophisticated ETL procedure is required. The ETL process is technically demanding and involves active engagement from various stakeholders, including developers, analysts, testers, and senior executives.

Extraction:

Data is extracted from the source system into the staging area in this stage of the ETL architecture. If any transformations are required, they are performed in the staging area so that the performance of the source system is not harmed. Rollback will be difficult if faulty data is transferred directly from the source into the Data warehouse database. Before moving extracted data into the Data warehouse, it may be validated in the staging area.

A data warehouse must be able to integrate systems with varying capabilities.

Hardware, Operating Systems, and Communication Protocols are all examples of database management systems. Legacy programs, such as mainframes, bespoke applications, point-of-contact devices, such as ATMs and call switches, text files, and other sources might be used.

Three Data Extraction methods:

1. Full Extraction
2. Partial Extraction- without update notification.
3. Partial Extraction- with an update notification

Extraction should not impair the performance or reaction time of the source systems, regardless of the method utilised. These are real production databases that serve as the source systems. Any sluggishness or lockup might have a negative impact on the company's bottom line.

Transformation

The data obtained from the source server is unusable in its current state. As a result, it must be cleaned, mapped, and changed. In reality, this is the critical phase when the ETL process adds value and alters data to create meaningful BI reports.

It is an essential ETL concept in which you apply a collection of functions on data that has been extracted. Direct move or pass-through data is data that does not require any transformation.

Loading

The final phase of the ETL process is to load data into the target data warehouse database. In a typical data warehouse, a large amount of data must be loaded in a short amount of time (nights). As a result, the load process should be adjusted for speed.

In the event of a load failure, recovery procedures should be set up to resume from the point of failure without compromising data integrity. Administrators of data warehouses must monitor, continue, and cancel loads based on server performance.

Types of Loading:

- Initial Load — populating all the Data Warehouse tables
- Incremental Load — applying ongoing changes when needed periodically.
- Full Refresh —erasing the contents of one or more tables and reloading with fresh data.

Steps for data cleaning

Monitor errors:

Keep track of the patterns that lead to the majority of your mistakes. This will make detecting and correcting inaccurate or faulty data much easier. If you're integrating other solutions with your fleet management software, keeping records is extremely vital so that your mistakes don't clutter up the work of other departments.

Standardise your process:

To assist limit the possibility of duplication, standardise the point of entry.

Validate data accuracy:

Validate the correctness of your data once you've cleaned up your current database. Investigate and invest in data-cleaning solutions that can be used in real-time. Some solutions even employ artificial intelligence (AI) or machine learning to improve accuracy testing.

Scrub for duplicate data:

To save time while examining data, look for duplication. Research and invest in alternative data cleaning solutions that can examine raw data in bulk and automate the process for you to prevent repeating data.

Analyse your data:

Use third-party sources to augment your data once it has been standardised, vetted, and cleansed for duplicates. Reliable third-party sources can collect data straight from first-party sites, clean it up, and assemble it for business intelligence and analytics.

Communicate with your team:

To encourage acceptance of the new technique, share the new standardised cleaning method with your staff. It's critical to maintain your data clean now that you've cleaned it up. Maintaining communication with your team will aid in the development and strengthening of customer segmentation and the sending of more focused information to consumers and prospects.

Big data validation protocols

Validation Based Protocol:

The optimistic concurrency control approach is another name for the validation phase. The transaction is carried out in three steps in the validation-based protocol:

Read phase

The transaction T is read and performed in this step. It is used to read the value of various data elements and save them in temporary local variables. It is capable of performing all write operations on temporary variables without affecting the database.

Validation phase

The temporary variable value will be checked against the real data to see if it violates serialisability at this step.

Write phase

If the transaction's validation is successful, the temporary results are saved to the database or system; otherwise, the transaction is rolled back.

Here each phase has the following different timestamps:

Start(Ti): It contains the commencement time of Ti's execution.

Validation (Ti): It contains the commencement time of Ti's execution.

Finish(Ti): It records the moment Ti completes its write phase.

- This protocol is used to calculate the transaction's timestamp for serialisation by utilising the time stamp of the validation phase, as this is the phase that decides whether the transaction will commit or rollback.
- As a result, TS(T) = validation (T).
- During the validation procedure, the serializability is determined. It's impossible to know ahead of time.
- It ensures a higher level of concurrency and fewer conflicts when processing the transaction.
- As a result, it contains transactions with fewer rollbacks.

Big data testing methodologies

You should anticipate finding that successful teams use the same kind of big data testing methodologies after analysing case study after case study.

Functional Testing

Data validation benefits from front-end application testing, including comparing actual results generated by the front-end application to predicted outcomes and obtaining insight into the application architecture and its many components.

Performance Testing

Big data automation enables you to evaluate performance under a variety of scenarios, such as testing the application with various types and volumes of data. One of the most significant big data testing methodologies is performance testing, which guarantees that the components involved have effective storage, processing, and retrieval capabilities for massive data sets.

Data Ingestion Testing

Include this form of testing in your data testing procedures to ensure that all data is appropriately retrieved and loaded into the big data app.

Data Processing Testing

Your big data testing plan should contain tests that focus on how ingested data is processed and evaluate whether or not the business logic is executed appropriately by comparing output files to input files.

Data Storage Testing

QA testers may utilise big data automation testing technologies to ensure that output data is properly loaded into the warehouse by comparing output data to warehouse data.

Data Migration Testing

When an application moves to a new server or technology changes, this form of large data software testing follows data testing best practices. Data migration testing ensures that data transfer from the old to the new system is completed with little downtime and no data loss.

Data processing and reporting issues

Big data processing collects methodologies or programming models for accessing massive amounts of data and extracting meaningful information for decision-making. The next sections go through some of the tools and methodologies available for big data analysis in a data centre.

High-Performance Techniques for Big Data Processing

Stages of data processing

Data collection:

The initial stage in data processing is data collection. Data is gathered from various sources, such as data lakes and data warehouses. It's critical that the data sources utilised be reliable and well-constructed so that the data collected (and later used as information) is of the best possible quality.

Data preparation:

After the data has been acquired, the data preparation stage begins. The step of data preparation, sometimes known as "pre-processing," is when raw data is cleaned up and structured in preparation for the next step of data processing. Raw data is thoroughly verified for mistakes during the preparation process. The goal of this stage is to a) get rid of poor data (redundant, incomplete, or inaccurate data) and b) get rid of bad data (redundant, incomplete, or inaccurate data).

Data input:

The clean data is then put into its intended destination (such as a CRM like Salesforce or a data warehouse like Redshift) and translated into a language that it can comprehend. The initial stage in the transformation of raw data into useable information is data entry.

Processing:

The data entered into the computer in the previous stage is processed for interpretation in this stage. Machine learning algorithms are used to process the data, albeit the method may vary significantly based on the data source (data lakes, social networks, linked devices, etc.) and the intended purpose (evaluating advertising trends, medical diagnosis via linked devices, etc.).

Data output/interpretation:

The output/interpretation step is where non-data scientists may finally use the data. It is translated, understandable, and frequently in the shape of graphs, videos, photos, plain text, and other formats. Members of the firm or institution can now use the data for their data analytics initiatives by self-serving it.

Data storage:

Data storage is the ultimate stage of data processing. After all of the data has been analysed, it is saved for future reference. While some information will be useful right now, most of it will be useful later. Furthermore, compliance with data protection law such as GDPR necessitates correctly maintained data. Data may be retrieved fast and simply by memb when it is appropriately stored.

Reporting issue

Issue Reporting allows a member to escalate a task when there is a problem. It also blocks it from being completed until the issue is resolved.

The top five issues come in big data

Finding the signal in the noise:

"There needs to be a detectable signal in the noise that you can identify" to properly exploit big data, "and sometimes there just isn't one."

"Sometimes we have to come back and say we just didn't measure this correctly or measured the incorrect factors because there's nothing we can identify here after we've done our intelligence on the data."

As a result, one of the most significant challenges firms confront when dealing with massive data is the traditional needle-in-a-haystack dilemma. The company added that big data appears like a hairball in its raw state and that a scientific approach to the data is required.

Data silos:

Data silos are the kryptonite of big data. They accomplish this by storing all of the beautiful data you've collected in distinct, independent pieces that have nothing to do with one another, resulting in no insights from the data because it isn't integrated.

The need to crunch figures to compile a monthly sales report is due to data silos. They're to blame for the sluggishness with which C-level decisions are made. They're the reason your sales and marketing departments are at odds. They're the reason your consumers are seeking a new place to do business because their demands aren't being satisfied, and a smaller, more agile firm is providing something better.

Inaccurate data:

Data silos are useless on an operational level, but they also provide fertile habitat for the most serious data issue: incorrect data.

According to a study by Experian Data Quality, 75% of companies feel their customer contact information is wrong. You might as well have no data if you have a database full of erroneous consumer information. What's the best strategy to deal with erroneous data? By linking your data, you may eliminate data silos.

Technology moves too fast:

Larger companies are more prone to fall prey to data silos for various reasons, including a preference to retain their databases on-premises and the slowness with which they make decisions about new technologies.

According to the Company analysis, like as telecommunications and utilities are experiencing "high levels of disruption from new rivals stepping in from other industries." In each of these industries, over 35% of respondents acknowledged this concern, compared to an overall average of around 25%."

Traditional players are, in essence, slower to accept technology developments and, as a result, are facing stiff competition from smaller businesses.

Lack of skilled workers:

According to the survey, 37% of organisations have problems obtaining experienced data analysts to help them make sense of their data. Their best chance is to build a centralised data analysis team for the organisation, either by retraining current employees or hiring new big data specialists.

You'll need to hire people that not only understand data from a scientific standpoint but also the business and its consumers and how their data insights affect them.

Configuration in Big data

It provides Hadoop Core configuration parameters, such as I/O parameters shared by HDFS and MapReduce. The configuration parameters for HDFS daemons, the NameNode, Secondary NameNode, and DataNodes are all included in the hdfs-site. xml file.

Slaves & Masters:

DataNode and TaskTracker servers require a list of hosts, one per line. The Masters include a list of hosts that must host secondary NameNode servers, one per line. The Hadoop daemon receives information about the Secondary NameNode location from the Master's file. Secondary Name Node servers are listed in the 'Masters' file on the Master server.

Hadoop-env.sh, core-ite.xml, hdfs-site.xml, mapred-site.xml, Masters, and Slaves may all be found in the Hadoop installation directory's conf directory.

The files in the extracted tar.gz file in the etc/hadoop/ directory are configuration files.

HADOOP-ENV.sh

It defines the Hadoop Daemon's (bin/hadoop) environment variables that impact the JDK. Given that the Hadoop framework is written in Java and uses JRE, one of the Hadoop Daemons' environments variables is $Java Home in Hadoop-env.sh.

CORE-SITE.XML

It is one of the most significant configuration files for a Hadoop cluster's runtime environment parameters. It tells Hadoop daemons where NAMENODE is located in the cluster. It also tells the Name Node the IP addresses and ports it should connect to.

HDFS-SITE.XML

It is one of the most significant configuration files for Hadoop's runtime environment settings. It comprises the NAMENODE, DATANODE, and SECONDARYNODE configuration options. It specifies the default block replication. When the file is generated, the actual number of replications can also be provided.

MAPRED-SITE.XML

It is one of the most significant configuration files for Hadoop's runtime environment settings. It includes MapReduce's setup options. By setting the MapReduce.framework.name variable in this file, we may give MapReduce a name.

Masters:

It is used to identify the Hadoop cluster's master nodes. It will tell Hadoop Daemon about the location of the SECONDARY NAMENODE.

On the Slave node, the Master File is empty.

Slave:

It's used to figure out which nodes in a Hadoop cluster are slaves. A list of hosts, one per line, may be found in the Slave file on the Master Node. Slave node IP addresses are stored in the Slave file on the slave server.

Optimisation changes

What is Big Data's role in optimisation?

Big Data is the process of gathering, consolidating, and analysing disparate data from many sources and forms to unearth new insights and create value. While Big Data originated in the consumer and financial services industries, the global industrial industry has recently been interested in its potential for uncovering important information. Big Data is, without a doubt, a critical component of Industry 4.0 and the Industrial Internet of Things (IIoT). Although Big Data can help with process optimisation, it's worth considering why it's just finding its way into the industrial world. Data, storage, and analytics are three crucial variables to consider. The cost of sensor technology and data storage has dropped dramatically in the industrial business over the last several years. These two advances have enabled manufacturers to capture and keep more data at a lower cost than previously possible.

So, how can Big Data help with optimisation in a normal manufacturing plant? Here are a few instances — only the tip of the iceberg when it comes to Big Data:

Nonobvious analytics:

Alerts and sirens have been pointing out obvious faults impacting plant operation for years. These simple warnings, which are triggered when a specific limitation is exceeded, have helped employees better their management of large and dangerous production settings. Alerts and alarms, on the other hand, are restricted to basic thresholds like HiHi and LoLo and the usage of equally basic operators (e.g. total change greater than X, value below the threshold for Y amount of Y amount of Y time, etc.).

Operational intelligence:

Historically, most practitioners would agree that their factories were run as a collection of interconnected silos. Each functional group – whether engineering, operations, or maintenance-concentrates on a single topic. Information and communication were mostly limited to members of a certain set of employees. Because of this constraint, few, if any, employees had a comprehensive operating view. The capacity of Big Data solutions to draw on information from across functional silos is one of its advantages. Consider how some diagnostic systems employ routine process data to detect typical mechanical defects that otherwise go undetected.

Changing mindsets:

Not long ago, data was only collected and maintained because a manufacturer was compelled to do so by law. Consider the rules enforced by government agencies such as the EPA, FDA, and NRC. At the time, there was a widespread belief that data collection was only a waste of money because it was done to meet some government obligation. With time and innovation, perspectives have certainly shifted. Many businesses increasingly regard their data as a source of revenue. One firm that has established a business around Big Data is General Electric, which has invested substantially in predictive analytics and saved clients millions in otherwise lost productivity.

Raising standards:

Production consistency is something that all process manufacturers aim for. However, even apparently, little elements such as the season of the year, the persons who staff each shift, and the source of manufacturing inputs, among other things, can affect consistency. Individually, they may not be significant, but when taken together, these minor variations can have a substantial impact on control. Staff members' capacity to detect and then adjust for so many tiny changes is restricted. Big Data can be of assistance. It enables firms to keep track of daily changes in production inputs, personnel, and environmental conditions, among other things, and to maintain – if not raise – their standards.

Some key methods for big data optimisation

If not properly standardised, big data is enormous, complicated, and prone to mistakes. If huge data isn't presented correctly, it can turn out to be erroneous in a variety of ways. Consider a name convention: Micheal Dixon can be spelled M. Dixon or Mike Dixon. Data duplication and skewed analytics findings are two issues that arise from an inconsistent format. Setting a standard format, for example, is an important aspect of big data optimisation because it ensures that petabytes of data have a consistent format and the potential to provide more accurate results.

Tune-up’ algorithms:

Implementing algorithms to analyse and fine-tune your big data isn't enough. The diagonal bundle technique, convergent parallel algorithms, and restricted memory bundle algorithm are some of the strategies used to optimise huge data. You must ensure that the algorithms are fine-tuned to meet the aims and objectives of your organisation. Data analytics algorithms are in charge of filtering through large amounts of data in order to achieve goals and give value.

Remove latency in processing:

The delay (measured in milliseconds) when retrieving data from databases is referred to as latency in processing. Data processing suffers from latency since it slows down the rate at which you receive results. Delays in processing are just unacceptable in an age where data analytics provides real-time information. Organisations should migrate away from traditional databases and toward cutting-edge technologies, such as in-memory computing, to drastically reduce processing time.

Identify and fix errors:

Fixing faulty data is an important element of big data optimisation. You can have the greatest analytics tools and fine-tuned algorithms, but it won't matter if the data is inaccurate. If your data is erroneous, you'll get false results, which can affect your ROI. In such circumstances, a data analyst will be required to go in and correct the data to ensure that everything is correct. Duplicate entries,

inconsistent formats, insufficient information, and even erroneous data can be found in big data. In these situations, data analysts must utilise various technologies, such as data DE-duplication tools, to find and correct problems.

Eliminate unnecessary data:

Because bloated data bogs down algorithms and reduces processing velocity, not every data collected is relevant to your organisation's goals. As a result, removing unneeded data is an important aspect of big data optimisation. When unneeded data is removed, the rate of data processing is increased, and big data is optimised.

Bringing it all together:

The key to accurate data analytics is big data optimisation. When data isn't correctly optimised, it causes many issues, including erroneous results and processing delays. There are, however, at least six alternative approaches to optimising huge data. These strategies include standardising format, tweaking algorithms, cutting-edge technology, correcting data mistakes, and eliminating processing lag. Data optimisation increases the speed with which data is processed and the accuracy with which outcomes are produced.

Report on evaluation outcomes and obtain sign off from required personnel

A report on evaluation outcomes is a written document that explains how the program was tracked and evaluated. It summarises the observations, conclusions, and recommendations from a specific assessment and suggestions on how evaluation outcomes can be used to enhance programs and make decisions. Though evaluation is a continuous process, the word "final" refers to the final report of a funding cycle or a particular evaluation task.

- The evaluation report should answer the questions "What," "How," and "Why It Matters" about your program.
- The "What" section explains the program's intent and activities about the desired outcomes.
- The "How" section discusses the process of implementing the program and offers details on whether it is performing as intended. The “How” (or process evaluation) is used in conjunction with performance and/or short-term result data to determine whether and why improvements were made during implementation.
- The "Why It Matters" question (also known as the "So What" question) explains why the initiative is important and how it will affect people. The ability to show that the program has made a difference is critical to its long-term success.

How do you write an evaluation report?

- Intended use and users: A discussion of the evaluation's intended use and users promotes clarity about the evaluation's purpose(s) and specifies who will access the results. It's critical to remind the audience in the assessment report what the intended use is and who the intended users are.
- Program description: The theory of change that drives the program is presented in a program overview. In addition to a narrative summary, this section often includes a logic model and a description of the program's stage of development.
- Evaluation focus: The reasoning and criteria for how the evaluation questions were prioritised are presented in this element, which records how the evaluation emphasis was narrowed.
- Data sources and methods: This section describes the assessment metrics, success measures, data sources, and procedures used in the evaluation. The transparency and credibility of assessment details can be assured by explaining how the evaluation was carried out.
- Results, conclusions, and interpretation: This segment explains how data was analysed and the collaborative method that was used to interpret the findings. This section goes beyond mere presentation by providing a meaningful analysis of the results. The analysis section is absent in several assessment reports, obliterating a crucial link between results and application.
- Use, dissemination, and sharing: This segment explains how the evaluation results will be used and how the conclusions will be disseminated. Clear, detailed plans for how the assessment will be used should be addressed from the start, as this will make the evaluation's course and communication of results much easier. This section should provide a general description of how the findings will be used and more specific details about the planned modes and methods for communicating the findings with stakeholders. This segment should also include plans to track distribution activities and a feedback loop for corrective action if necessary. The dissemination plan is an integral part of the assessment plan and study that is often overlooked.

Methods to communicate final sample results:

Meeting: This is a personal, interactive exchange often succeeded by a written follow-up to communicate changes in compliance management strategies.

Video conference: These allow people in different locations to hold interactive meetings to discuss the changes in strategies in real-time.

Telephone conference: These enable participants in different locations to share information.

E-mail: This is an instantaneous medium for formal notices and updates, as well as informal exchanges.

Report: This is the official documentation of the activities of any department or organisation.

Presentation: This method usually comprises a formal proposal, update, recommendation, or report involving audiovisual material, slideshows, and statistics.

Forum: This method allows members to post information publicly and efficiently in a centralised location.

Obtain sign off from the required personnel

It's now time to finalise the reports termination and seek sign-off on the project's final report. By ensuring that the results follow all of the criteria specified in the scope statement, this process ensures that the consumer and stakeholders have officially approved the product. Producing and distributing a final progress report for the project that includes approval signatures is one way to get sign-off. The final status report differs from the previous periodic status reports you've been making.

The project's priorities should be summarised in the final report and the main achievements and deliverables that have been approved and completed.

The aim is to provide a high-level overview of the project, including its achievements and challenges and the dates when stakeholders approved the main project deliverables. Keep it honest, but not brutally honest, since this report is for public consumption. Consider preparing a confidential report or briefing to send to the project sponsor or relevant senior managers if significant issues arose on the project due to any stakeholders or team members that were especially difficult to work with. Put any other political hot potatoes that emerge during the project in the confidential report if you don't think they should be included publicly in this status report.

Arrange for walkthroughs.

If you schedule walkthrough sessions, you'll have a much better chance of getting approval. It will also allow you to double-check your understanding and make changes if needed to prevent delays. It's cool if the stakeholder believes they don't need a walkthrough; however, they should be provided one.

Request signatures via email with a deadline date.

Often request a signature via email. It is not appropriate to treat a lack of response as an implied sign-off.

Have a due date in the email and follow up if the deadline is missed to learn why. If it's due to a lack of time, inquire as to when they should meet. Provide a second walkthrough as well.

Allow the reviewer to indicate whether or not they would sign off on the document based on the input they provided being presented. Often, inquire as to whether they can sign off on the document without having to review it again or whether they must first review the changes that have been made.

The process ends with a formal notice of approval to the stakeholders, clients, and project sponsor after the stakeholders sign the acceptance document. The project manager is in charge of disseminating this final piece of information. According to the communications schedule, this notification should be prepared, dated, and sent. If the project is only for internal use, an e-mail notification might suffice. If the project was finished on time and budget, I suggest using the old-fashioned method of sending a formal note via the mail. This is the last time I'm going to say it: The approval paper should be kept in the project notebook.

Caveats to be aware of

If it would make it easier for them to determine if sign off will be easier if they have caveats in their sign off, they should do so. If a stakeholder has concerns that cannot be managed or resolved until later in the project, they can feel more at ease signing off with the caveats they are concerned about.

Protocols and Procedure for big data to write scripts and queries

To various individuals, big data implies various things. For me, it's all about enabling the right decisions to solve business challenges by often correlating disparate and complex data to key business levers in its purest form. Those who grasp the levers they have to increase performance are the most effective corporate leaders. Big data translated to such levers can help improve decision-making and result in genuine performance gains.

Collect

The first step appears straightforward, but there's a catch: while gathering and combining data, go beyond your local data sources and needs. By definition, big data is as thorough as you can make it. The collection stage will be made or broken by cross-functional awareness of features and capabilities. Add external data to the view you've created with your internal data sources.

Validate

The raw data should be consistent and thorough. Companies frequently exploit skewed data in the mistaken belief that analysis would cover any flaws. At this stage, strong project management is required to verify that the data is accurate and up to the job. Invest in human capital, not just technology, as a first step in addressing big data.

Analyse

Before looking at the facts, consider the final aim. A smart manager or consultant will be able to provide you with more than "artificial intelligence," which is merely data that has been organised. Although it may appear impressive in chart or graph form, it frequently lacks context.

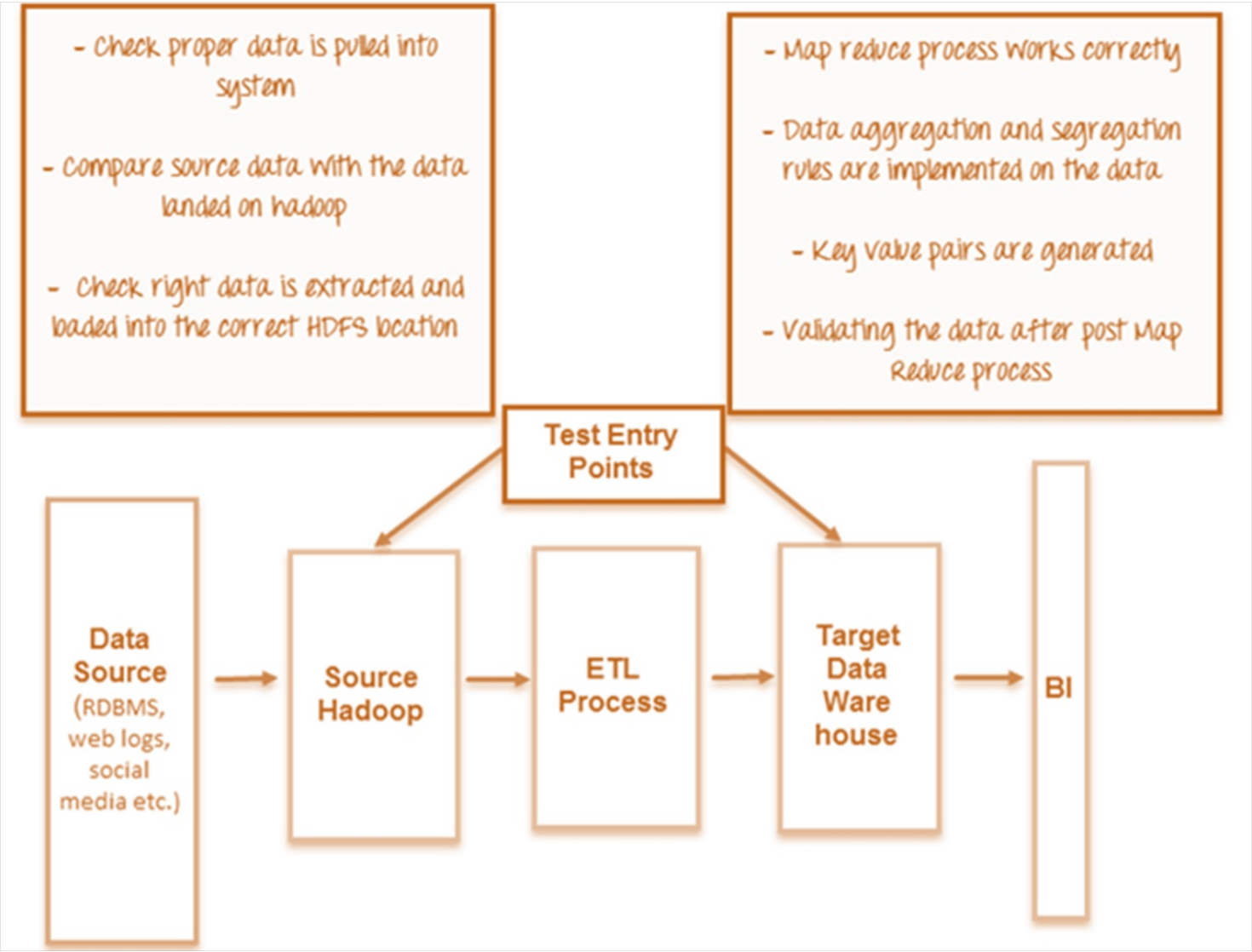
The ultimate test of data analysis is how fast and easily it can be completed. The findings from the analysis step will not be consistently applicable or useful if you struggle to draw parallels across groups of data or if obtaining raw data is challenging.

Find Your Golden Thread

To give the performance insights needed to manage results, several data sources are frequently integrated. The common connection becomes the golden thread - a shared characteristic that, once identified, connects enormous volumes of data in a way that allows your management team to achieve goals they couldn't before.

Structured procedure and extensive data analytics are required to find the golden thread in a sea of data. It necessitates a thorough grasp of internal data sources and the supplementation of external data to create a full picture.

How do you write big data test cases?



Big Data Testing can be categorized into three stages:

Step 1: Data Staging Validation

The first stage of big data testing, also known as Pre-Hadoop, consists of process validation.

- Data validation is critical so that data collected from various sources such as RDBMS, weblogs, and so on are verified and then added to the system.
- To ensure data consistency, compare the source data with the data added to the Hadoop system.
- Ensure that the correct data is extracted and loaded into the correct HDFS location.

Step 2: “Map Reduce” Validation

The second stage is the validation of "Map Reduce." The tester performs business logic validation on each node. Following authentication, they are run against multiple nodes to ensure that the:

- The Map-Reduce process is flawless.
- Data aggregation or segregation rules are imposed on the data.
- The ability to create key-value pairs is available.
- Data validation is completed following the Map-Reduce process.

Step 3: Output Validation Phase

The final or third stage of big data testing is the output validation process. The output data files have been created and are ready to be moved to an EDW (Enterprise Data Warehouse) or any other system as needed. The third stage included the following:

- Ensure that the transformation rules are correctly applied.
- The target system must ensure that data is loaded successfully and that data integrity is maintained.
- It is ensured that there is no data corruption by comparing the target data to the HDFS file system data.

Lecture 5 - XBD401 T3 20230823 1633 XBD402 Lecture

All right, well, welcome Lisa and Thomas.
To this session here, so this is a long one,
because this is the start of the second unit,
as you can see here.
So we have spent three weeks in 401.
And 401 was just looking at,
you know, where do data come from?
So we did a lot of things around sourcing for data.
But now comes the next stage of it, where 402 is really about the title of this unit is testing big data samples.
It's a bit misleading having that word testing,
but what I would probably say to make it simpler is the fact that we are now in the unit to help us to look at
Preparing the data that we have source for analysis.
So this is pre analysis.
All right, so some people call it validation stage,
because this is where you actually check your data and make sure that all nice and clean.
Make sure that's validated so that we can then run the analysis on it.
So it's a really important step.
I mean, I have to say maybe about 20 years ago,
or maybe 30 years ago now,
when I first started this business of analytics,
they didn't call it analytics in those days.
It was just called number crunching, you know.
And this was something that we didn't do a lot in those days,
because we weren't, you know, this discipline just hasn't started.
But now if you're a data analyst, if you're dealing with numbers or some kind of a data,
this is a really, really important stage.
Once you know where to source for data,
and you're starting to gather them,
this is where it happens. This is where you start to go in and clean,
get rid of duplications, get rid of empty, you know, empty cells.
What happens? You have data that's just missing.
These are really important things.
So I'm going to spend a fair bit of time to this evening,
just going through those terminology,
and we'll be looking at some of the exercises,
which I hope would be helpful.
So I'll be looking at both Excel,
and I know you guys love Excel.
And I'll be using also a little bit of Python to help us to handle these sort of situations.
So to me, this is the best practice sort of a discussion,
or a narrative around data.
The next just to give you some pream as to what's going to happen from here.
403 is about the actual analysis itself.
So once the data is all been nice and clean,
we are now going to do the analysis on it.
So this is where we start to, you know,
go back to our statistical background, our roots,
and make sure that we know how to do modeling and those sort of things.
So we're going to talk a little bit more statistical, I guess,

in a sense in 403.
And in 406, which we're going to wrap up this particular study skill set,
is 406 is actually about visualization.
So it's about plots and graphs,
which will be useful for almost anyone nowadays, you know,
to draw plots, whether it's pie chart or histogram or whatever it is.
We'll be using tools again.
Hopefully we'll be using some of the tools that I can make sure that you know,
the students can actually download easily.
We're not going to make make you buy tools and so on and so forth.
So I have to look for tools that are more or less easy to download.
And you know, there's no fees to it.
Okay, so that's that's the plan.
I just wanted to say a little bit on the study plan.
So let me see, I think I have it up here somewhere.
There it is, okay, just want to say a little bit about it.
I had a couple of questions about, oh, you know,
I kind of missed the deadline.
Can I ask for extension?
So we just finished 401.
All right.
And there's two assessments in there, 81 and 82.
So the due dates for both of them, but a 22nd of August, that was yesterday.
So that created a bit of, I guess, urgency, anxiety for some students.
But what I'm saying is don't worry too much about that.
Worry about it to just to a certain extent that it's at the back of your mind.
You know that it's, it's due.
And just do the best you can to finish it off and submit it.
Because as you can see, we're just going to go through, you know, 402.
Again, there's another two assessments here.
403, it, you know, the more and more assessments coming down the,
down the park line.
So it's just a bit of a warning, not to leave things to us.
And because you're probably getting anxiety attack because it's just so much work to do.
:
So try your best to, to, to keep on on track.
I know it's not easy.
We're all busy.
We have full time jobs and we have kits and we got obligations and bigger football games.
You know, that sort of things.
We just need to be a best, I guess, consistent in our, you know, just try to spend at least a couple of minutes.
So just put it in the evening or every three days or so.
And just knock a little bit of this off.
As I said in your assessments, I don't require you to write me essays.
It's just probably one or two lying sentences that will help me to,
to sort of mark you as yes, you know, you understand what you're, you're talking about.
Alright, so that's just sort of a reminder.
For 402, which is about, you know, the call it testing big data samples.
This is the official name of the unit when I tend to call it pre analysis or staging or validation.
I think those are more appropriate words to, to describe what's going to be happening in this particular unit

And the due date for that is the 12th of the following month.
Okay, so that is the you have access to this particular study plan.
So just keep it close to, to, to, to yourself.
Alright, and also I made a change here in, in 2024, this skill set is expected to be offered in term one and
We have decided because of the, it is, it's quite, it's getting more popular.
We've decided to just offer it right across term one to entry.
Alright, so it's every single term except for term four where I can at least go on my holidays.
Alright, so that's the, that's the plan.
Okay.
So we have, we have, we got Eric and Glenda on board now. Welcome. Welcome.
Let's see what was I going to look at.
So I'm going to cover a couple things I got Excel open. I've got a spider open because we're just going to play with a quite a complex data set.
But you know, it looks horrible. It looks really hard to understand, but we'll, we'll take some time just to get
And you know, to be fair, there will be one fairly simple question in the first assessment for this particular unit for zero two.
So watch out for that.
I think we'll start here.
Yep. So what I'm going to go is just to talk through this, not all of them, because this is the, the, the election notes for this particular unit.
And just to remind those of you, where to get to them, it's true. Obviously, your, your, your, your learn line side, but under 402, if you click on that, you open it up and towards the bottom.
And you'll see this link for zero to lecture. You can click on it. And that's where you'll find the, the PDF version of the PowerPoint slides.
At the end of today, I will also add on to the bottom of the lecture, the, the first assessment, 81, those of you who wants to get ahead and there are a few of you who are ahead of the curve.
It's just that I didn't put it there now because I just wanted to, to figure out how far we would go this evening without stretching the clock. And you know, what, what kind of questions I can put in there.
Just one question, but multiple parts, it should be pretty straightforward. But the whole idea is for you to, to just go through some of these syntax and some of the commands, the new commands in Python.
Okay, that's the plan.
So as you can see here, 402, just again, is you can go through the chapters again. If you look at the headings of the chapters one, it's all about validation.
Chapter two is again about validation output. Chapter three, chapter four is about, again, optimizing, but it's really about validation.
So as you go through this modules, I guess I can call them, you'll find that it's a really good set of practices.
But I have to say, I have to say that these modules here are very, I find them quite what's the word for it to
So it's a lot of words and narrative around models and so on, which it's, which doesn't really quite reflect the real world now. So I decided to, to, to look at a more realistic view by looking at examples and stuff like that.
Okay, let's, let's get into it.
I'm going to use these slides here. Hopefully you can pull them. Any questions so far?
So Ingrid just join us. That's good. She was telling me that she was coming in.
Alright, so I assume that you are hearing me loud and clear. I'm looking at my recording. It's, it's on and the, my mics on. So everything's ready to go.
So as you can see over here, I can make it a little bit bigger.
You can see that there's a total of 70, 79 slides. Now I'm not going to go through all 79 slides. That's going to be a, that's, yeah, that's just too much.
But I'm just going to focus a little bit on the slides that I actually develop and put it up here.
So it's going to be the first few here. As you can see, it's quite a few that we need to go through. But from slight 21 onwards, which starts from what is big data.

I'm going to leave that to you to read. Alright, from this point onwards, it gets kind of it's easy to understand so you can read about what it is.
It's more like a review. It talks about sources of where big data comes from. I'll just let you read through that. And if there any sort of slides that's that's hard to understand or problematic, we can always talk about it Okay, so see it talks about sort of types of validation. You know, we don't really talk about this sort of types of validation now.
Alright, let's start from the very top.
And along the way, I'm going to, I'll probably have to pull in Excel as, as, as just as an instrument to sort of go through some of these things.
So before I forget, let me acknowledge all First Nations people across the land on which we live and work. So I'm in Darwin, some of you may be in analysis elsewhere.
And we pay our respects to elders, past, present and emerging.
Okay, back to something which I've talked about probably too many times now, but it's really important that I just want to just wrap this up because this is really the foundation of of the different components.
And why we're doing cleaning. So I'm going to use the word cleaning. It's about cleansing, cleansing, a data, getting rid of the noise, getting rid of what we call a dirty data, making sure there's no missing data.
Making sure that the numbers are within the range.
Alright, an acceptable range. So this is just a, probably a very holistic view at the EW, which is the enterprise data warehouse.
And it's good to just review this again so that we can get to an understanding.
This is the foundation. This is the basic, the basic off of what we call data analytics or data science.
So what we have is the intent to create an enterprise data warehouse.
And from there, you can have a whole bunch of tools trying to produce reports out of it.
Some basic analysis and so on and so forth.
But we really need to start the discussion from the left hand side, which is the data sources, which you know by now can be from disparate various very different kind of data sources.
Some of these are structured data, some of them are unstructured.
Alright, so this is where the raw data originates from.
It could be from systems within our organizations, our enterprise resource planning systems, our finance systems, our customer relationship management systems.
We can also have individual SQL is is is is is a query language actually is database.
So we have our own database, which may be storing very specific sort of information and data.
Another source could be idle tea. This is going to be huge.
With a lot of devices now having the ability to produce and collect and to disperse and to disseminate data.
This becomes huge. I know in my in my world, which is in the healthcare in the healthcare industry.
We are having more and more medical devices out there in hospitals and also non clinical environments where these devices, which start to collect information and just keep on pumping the information out.
Why this lead through service so these service will be collecting roles and roles of transactional data.
Alright, that's going to be huge and think about this in your own space.
We also have you know data from social media and that again is huge from Facebook, Instagram and so on.
And that is truly a very unstructured sort of a data that's been collected even from websites.
And of course, traditionally what we call the you know, these are more like very quick structured data like You know tables, rules and columns, spreadsheets. Those are more more towards like the structured sort of So and what we have here is the movement towards more of a storage layer.
So this is where if you see this this this symbol here, that's that's the data warehouse that we're trying to Alright, so organizations are hitting towards this area or are in this particular state at this time.
They can have a number of data warehouse, but in most cases what we have is a data warehouse.
And then from there we're going to have data marks. I'm going to explain that in a minute.
So the intent is moving from left to right. Now there are a couple of things that can happen at this stage.

This layer here between the sources and storage or some of you may call it the database is an area which is often talked about.
And usually referred to as a ingestion layer. Ingestion comes to the word ingest.
I and GST ingest means to I guess in English means to ingest to ingest food, for example, to absorb food to absorb information.
So I think those are quite those are the it's quite synonymous to those terms. So ingest is to is to feed is to
And I guess that makes a lot of sense because it's about pulling data out of sources from here and delivering it to a warehouse ingestion.
So what we have is for some organization they tend to have a staging area.
A staging area is sort of a meat area between this and that where they have something that they can play
And before I'm making sure that this is all nice and clean before pumping it over to a data warehouse.
This is also this is often known as the staging area. Now in terms of the the extraction.
The transformation and loading. That's where the difference is just to make it a little bit more official.
And it's like now most of the time we head from here from the sources right into the data warehouse.
We still need to extract. But the terminology sort of changes.
If you go directly from the sources to a storage layer. It's known as E L T as opposed to ETL.
See, that's a difference between these two sort of.
It's just a switch of those things. And the switch is based on the fact that if you do have a staging area within your organization, whether it's bipolarcy or by, you know, by that's your common practice.
If you do have a staging area, most of the time you refer that as a ETL which is extract, transform and load.
But if you don't have the staging area, it becomes a L T which is extract loading and transform.
All right. Now I know the difference is kind of silly, but just to let you know that is a difference between these
So the distinction between ETL and E L T approaches is in the order of events.
Okay. In ETL, the transformation happens in a staging area.
In a more modern approach, the ELT conducts all the transformation jobs even within the warehouse.
Okay. I think that's as much as I want to say about that.
So over here, when we talk about cleansing, which is this layer in the middle here, moving from data sources to your data warehouse, we have a lot of things that happens.
We, we have to ensure that the data get clean.
We need to get rid of duplications, which is also known as de duplication.
We, we need to split some of those data. We need to join data tables together.
So fair bit of things to do within that space, but within the storage layer, once we are happy and once we start to get all these all these records into our data warehouse.
This is where we start to store the really important information.
So as mentioned before, the data warehouses are mostly relational databases, meaning roles and columns.
All right. Within this space here, there's also a database management system.
So you may have heard of those DVMS examples of DVMS are my SQL oracle, the one that I'm familiar with, which is IBM DB2 and Amazon RDS.
Those are some of those database management systems that is often used within this storage layer.
Not up here is metadata.
Metadata is simply is data about data.
All right, data about data.
So it's stored usually as a separate module within the storage layer.
And it's usually managed by a metadata manager.
Now, these data marks here, what are they useful?
Now, in most cases, and I know in anti-help, they do this. We have a huge data warehouse where we start to think about it.
And how do we split that up in a logical way? All right, this is a logical diagram.
Now, we need to in some cases, we need to have smaller subsections of our data warehouse.
Now, each of these subsections may be pointing to a particular subject area.

So in a business environment, this one up here could be functional in terms of this could be sales data.
Very specific to sales data.
And most of the time used by our sales people, this one at the bottom could be a finance data.
And the one below could be maybe business development type data.
So what happens here is a separation into various data maps for very specific purposes.
So as I said, this could be sales, this could be finance, this could be for marketing.
It also helps in terms of security because now all the sudden you have data maps which are actually very Data sets used by a very specific, authorized set of users.
Okay, so that's that's the sort of an intro to what data marks.
So it actually created from a data warehouse.
Now at this stage, are there any questions that you have before I kind of move on?
Hopefully that was that was pain painless.
No. Okay, good.
Now the last layer which is the presentation layer is where the analysis comes in.
So we have a bunch of BI tools, BI stands for business intelligence.
So in in anti-health, the group that I started working with in Alice, they were they call themselves the BIT.
The business intelligence team. So, you know, go back to the organization, us around and you know what, what's the name of the groups that are involved in this sort of analysis.
So we have a bunch of tools that we can use like Tableau and, you know, power builder and in the, what's the I want to call, there's a few of them, cognos and so on.
So all tools that we can use. There's also very specific reporting tools, crystal used to be traditional crystal Even some of our business applications like, you know, maybe systems could be pulling information out of the data marks or the data warehouse.
Now the way that they do that is true these things here so they can actually make a typical SQL query.
Alright, so at this stage, we haven't really talked about SQL and it's not the intent of this unit to talk about SQL, but it's actually a, it has its own language in terms of trying to make trying to find information such as, you know, can you extract certain columns out of a particular data set.
So I just want to look at the name of all the sales people along with the sales contribution, for example, so it's just a query language.
That's what a QL stands for a query QUERY kind of a language.
Or you can use API. So there's some applications that provides people with APIs, application programming interface. I think that's what it's called.
Like Facebook, for example, you can they they provide APIs for people like you and me so that we can use those APIs which I just simply scripts or software that we can pull out certain information from from Facebook because they've given us the connection into those those those data sets within Facebook that we And so that's a known as those are known as as APIs. Okay, so that's sort of a really rough view as to as to the different components and I know I talked about this the last few weeks.
Hopefully by now I'm just trying to just sort of consolidate this this whole thing all together and one one Now there are attributes for these data warehouse.
So they do have a likeness of the original data source surprise surprise because they are actually extracted from a original data source that they should have a very similar likeness to the original data.
Now the data stored in a enterprise data warehouse. I just used the word data warehouse is always standardized instruction.
Alright, it's always that the whole reason behind that is just just just a possibility or the easier possibility for end users, those people who are making inquiries or trying to produce reports to make those reports easier.
So they know how to get into the the data sets and pull out information is all pretty much standard and very structured in terms of the process itself.
These data warehouses are also very subject oriented. So this is where we start to talk about the data marks because the data marks are typical data models.

So you can have a sales data model. You can have a marketing data model where the columns are just quite related to sales or finance or to marketing very subject oriented.
These data warehouses are also quite time dependent. So they're basically historical data way in the past. They're not here to talk about what's in the future.
Alright, very time dependent. Usually historical data and they also quite non wall towel in other words once placed in the warehouse, the data is never deleted from it.
Okay, so you can manipulate the data you can modify and update it due to sources change in the sources, but you're never meant to erase from the data warehouse.
Again, deletions are actually very counterproductive for analytical purposes. So another reason why it's non wall towel is because of analytical purposes.
Alright, so these are some of the attributes or characteristics of a data warehouse.
So let's look at some of these different, I call them tiers to explain is a little bit more easier. So this is a tier This is a tier one is really basic kind of an architect architecture. So again, this will make complete sense to I hope on left hand side are all the different sort of the whole variety of data sources, whether they're structured or unstructured, they're then pulled in together, whether through ETL or ELT, whether you have staging or not, you get you pull them together, you integrate them and you put them in the warehouse. Okay, like so. And then you have these people out here who are trying to make inquiries or queries and trying to produce reports out of these data.
Alright, that's tier one simple straightforward and that's where most organizations would would start.
And then what we do is we, as you get more sophistication, more understanding of the concept of how important data is and the fact that you have different departments coming in asking for their own very So we have like we have in for example, in anti health again, we have very specific hospitals and within the hospitals, we have two main hospitals in an anti obviously in Darwin and Melis for each one of the hospitals, we have the different departments and they are, you know, they're concerned is pretty much around their own sort of data. So for example, the emergency or ED only be pretty much focusing around ED sort of So if you look at this, this sort of a diagram here, you can see the only difference here between the first tier and the second tier is now we have the the data map level here.
So where we start to look at a data warehouse and have sub sections of it.
And we call them data, so we have data marks that are quite related to sales and we have marketing and we have HR related and of course we have project management kind of data.
And then each one of the users out here would actually make inquiries against a you know a data map of choice of interest.
Now we can start to break that up into the next layer, which is a treat here layer and architecture. Now why am I talking about all this because it sounds a bit technical, but when it comes down to designing this, it becomes really important because even though you're a data analyst and you say all I do is crunch numbers, you need to know the kind of sort of a sort of an infrastructure view and architectural view of of how actually the data is split in your organization.
So over here we have a situation, which is a little bit more complicated.
So we have yes, we have our data marks here, but now we're going to produce cubes and I'll explain what cubes are and what we do here is is to actually empower the users.
This is where we will then empower the users and I'm trying to push this through anti hell and beginning to Because what's important now is you're going to have cubes created out of sales and again, don't worry too much. I will explain what cubes are just look at it as a cube, a rubric cube, where information is is actually A cube is like a little, you know, has sites and all that so it looks a bit like this. I'm just going to flash this quickly, but I'm not going to talk about it until after this. So that's a cube.
All right, that's a cube.
So what we need to do is create a cube for each one of the sales, for example, one for marketing, one for project management, and we're going to actually pass this and send it through email if you can to the end So that they can play around with these all app cubes and then produce their own reports based on their

And the whole idea behind this is really important and I've been trying to preach this for quite a number of years now is to empower the users because now our bi teams, our analysts are totally flooded with requests You know, the minister wants this, the department hit ones that the CEO wants that, you know, it just comes in every day, just amazing. It's just just too much.
So this is where I started back in, I started talking about this when I was in the Middle East, when again in a university environment, we also, we have a lot of data marks.
Obviously, and then we created a cubes from student services and from there we passed the cubes down to all the different divisions to the faculties and so on so that they can produce their own reports on, let's say, student grades and all that.
So this is where the big line is crossed, this is where we actually empowered the users because now the current situation is guess who's doing all the work.
People like you and me, all right, the analysts.
So now what does this cube all about?
The cubes have been around for a while. It's nothing new. All right, it's been around for quite quite a while.
What we are quite common to is the two by the two dimensional cubes.
Over here, we have more than two dimensions. We have three dimensions.
All right, so what we have is we have information.
So you can look at this information in terms of like columns. So this is time.
We have source. So what source in terms of the country.
And we have routes. So obviously this all app is about trying to look at the number of packages sent and when was it sent between different sources, which are countries.
So at what time?
It's the first half of the year, second half of the year, into split in the quarters.
So this is where we start to have a cascading view of time.
So we start to break time into into first year, first half of the year, second half of the year.
And then we split that into q1, q2, q3 and q4. Some organizations may then split that into weeks.
So we have one, two, three, four, five all the way to week 52.
All right, and that's just on one dimension that is time.
We also have sources in this case, countries, countries in the Northern hemisphere, Southern hemisphere, Eastern, Western by continents over here.
You know, there are you call the shots as a designer of cubes.
You will then decide, you know, how far do we want to go?
Is to go in and drill down. You can drill down and you can drill up in terms of routes.
All right, so where is it going? Is it going by ground or is it by air?
So on and so forth. But the actual values in each one of the, of these dimensions is actually the packages.
So how many packages were sent?
All right, between Asia, let's say in Europe in the first quarter of 2022.
Using air.
Plains. All right, so you can, you can address hundreds and thousands of different combinations of questions that users will throw at you.
All right, so just imagine if, if these are all thrown at you from a manager who wants to know all these
You know, so it's just, it just takes too long.
So give them a cube.
Teach them how to use the cube because it can be fairly simple is true Excel.
And they can then produce their own reports based on their own questions.
All right, so I guess at this point, let me just give you a very simple example of, of these cubes says.
Let me just get down to Excel.
Now, has anyone use pivot tables before?
In Excel.
Okay, blender, pivot tables, okay.

All right, this is, I'll show you an example of a pivot table in Excel.
And tell me whether this makes sense to you. This is how this is a very simple.
It's probably too dimensional at this stage.
You know, dimensions you could, you could reach out to.
Obviously, more than 10 dimensions that she, as you can imagine, that gets very complex.
But what we have here in front of you is a is a, is a lovely Excel spreadsheet.
All right.
It has to do with products that's been sold.
It gives you the name of the customers that have probably bought these different products in the year.
So this is here. It tells you that this, this year, if you're able to go down, it's, it's quite a bit of information
Let's have a look at it.
I think it's 900, whatever.
900 plus, yep, 950 records.
It's not large, but it's still sizable.
We have a name of customers in the year. So it's all 2013.
So the quarter is Q1234 and the amounts that was purchased.
All right. Very simple table and in the timeline.
Okay.
Typical sort of a very structured data set.
But you know, you're going to have questions from different people asking you different, you know.
Can you list down the top three customers?
Can you tell me how much we are making each quarter selling Alice.
You can think of all these different combinations and you drive you crazy because that's what people want to
And that's what managers want to know.
And because they can't do pivoting, for example, they will come and talk to a typical data analyst.
So what we do is we have this thing called the pivot table.
And this is just the most basic form of all app.
We talked about all app didn't we?
Online analytical.
This is where you start to start to hopefully look at all that if I can remember how to do this.
All right. So I can click on any area here. So this setup.
This is just this is more of a in in a typical Excel.
This will be blank. But this is this is how we set it up as something very, very simple.
And but more important is the stuff that goes that looks to the right hand side.
My mouse is a bit naughty today.
Oops, I can. That's pretty useless.
That doesn't enlarge my this.
Okay, never mind.
So what we saw up here and as a data source was five six columns. Right.
So.
So it knows. So when you start to build this pivot table, it knows that there are six of these one, two, three, four, five, six variables or six headings or six labels.
And let's say we try to produce a report. So this is given this all that is, is let's say we is given to a manager.
And manager learns how to do this. So the manager wants to know, for example,
let's mix wants to know products. All right.
And wants to have an idea of products as a role, but roles will be products.
All right. I drag and drop and all the sudden you see all my products appearing on the left hand side.
And for the columns, let's say I want to know how much did I sell for quarter.
So I drag. I highlight the quarter and I drag it down to the column.
And all the sudden you can see this things trying to to to take shape.

It still doesn't look too pretty because it's still it's not giving me the kind of results that I want.
I want to obviously to look at amounts because that's my value.
Amount spent by my consumers, customers, I drag down to my values.
And all of a sudden it starts to make sense now. All right. So I have products as a row.
And now my quarters Q one to Q four. It even adds up for me the totals that I've spent is all within the year
I can now be a little bit more fancy by saying I want to filter.
All right. I want to filter by customers.
So I want to know, for example, my customers. So I drag my customers down to my filters.
And there, oh, there's a really daughter in here. Cancel. Where do I do it?
Oh.
When I clicked that, what happened was it actually created this thing here.
It created the products and within the products, it sort of listed my customers.
You can do that. I was hoping to to drag this down to my filters so I can.
It's really there.
Yeah. Okay.
So what we have is we have customers now in my filter.
And if you see up here, that's where my customers are.
And I can use this drop down box to say, you know, I want to look at Anton because I like Anton. He's my
I want to find out exactly how much he spent on my products. I can click on Anton and say, okay.
And over southern this is this table sat here, this table here within Excel is giving me information related to
I'm giving me a list of all the products and of course, it looks like he hasn't bought anything in Q1 and Q1 and
So that's something that I might sort of put at the back of my mind, for example, if I will say a product manager or sales manager, it's interesting. Why is that so?
All right. But at least I get some information over here.
That's, that's just an example of what I was trying to talk about here.
In terms of an online analytical and all that.
So the pivot tables are very, very simply, very simplified, all that, all that.
As you can imagine now, this will start to make sense because all of a sudden now that you have seen the pivot table, which was actually here in this form.
Okay. So what you can do is you can give this to a user and they can just drive this on your own.
Okay. Any questions there?
Thomas, you didn't, you said earlier that you didn't see this now.
Would you, would you see that's a fair bit of value in this within your organization, for example?
Yeah, there's also an opportunity to use it as opposed, I'm not directly related to much data, but more pro project delivery, but it's definitely good to be able to factor in the information to use in the future.
Yeah, yeah. And it gives you the ability to drill up and down.
I think that's a beauty behind what all that is.
All right. So I just want to give you a simple example saying that yes, you can do analytics, even using Excel.
This is where actually, you know, a lot of people would would tend to look at it that way.
They see themselves as as as analysts, but they're going to be there just very specialized in using Excel.
All right. They may not be programming in Python and so on, but they're really good in doing this sort of stuff
All right. Thanks for that, Thomas.
This one here is just a something that I stole from somewhere, where it sort of differentiates between data warehouses, data lakes and data mods.
Now data lakes is something that I'd have to not talked about, but I did talk about data warehouses and I did talk about data mods.
So we know that data mods, the whole I guess the impetus or the intent is to actually be very specific about the needs of very specific departments, for example, marketing sales, finance and so on.
And you create that from a data warehouse.
All right. Data warehouse on the other hand is just looking at the needs of the entire organization.

So it gets a little bit different in terms of its purpose. In terms of data lakes, the reporting need for this sort of for this sort of structure is really for big data.
As we get into the world of big data, this middle column here, sticks on a little bit more level of importance. All right. It's you create data lakes because we have very huge amount of data are in there, whether it's structured or instruction, you know, we can we can still use data lakes and we can have very sophisticated tools that would make inquiries again.
These data lakes for reports and so on.
So I'll let you go through this. It has the different sort of looking at some of the features, such as the data type of data that store between the three different types.
The sources were that internal external.
The size is something that I would tend to question because this has I think this has changed organizations I guess policies and procedures around the size. Some of them may not even have a size as a measure of whether it becomes a data.
A lake or a data mod. These are just sort of very generalized guidelines. You'll be interesting to find out from you guys or from from from you know, we certainly do not have this in in in empty hell, but again, you know, more of the bigger I guess the big bigger players in in the world in Australia would would actually have something like that that would differentiate between each one of these different structures.
So just keep note that that away house difficult to set up and this is so true because to get to the stage where you have a really good robust well defined well running if efficient effective data warehouse takes quite a while different levels of you know just just levels of revisions.
And then once you get that done, obviously the data months are easy to set up that that would be a logical conclusion that data lakes themselves are quite difficult to set up.
Just a little nice table that I thought I would share with you guys. This is just another picture.
It comes up comparing the data warehouse with a with a data marks here. So data marks are very domain specific or subject specific such as sales marketing finance.
Again, looking at your industry, hopefully those data marks are well defined as different functional areas. You can have boundaries between them and then we have data lakes over here.
All right, which are larger amount of information, structured unstructured so on and so forth and look at the users, they're kind of different.
They're most sophisticated rather than just a normal B.I. user who could be from your department. These are actual specialists.
Okay, so that's the kind of differences. Just to wrap this discussion up the data warehouse technologies.
In terms of where do you store them now do you have on sites on premise sort of servers.
As opposed to cloud cloud technologies are becoming much more popular now.
Definitely so there's a term that you will probably see quite often which is the AAS which stands for data warehousing as a service, which means that they are service providers out there in the world and there are some popular ones at the bottom here.
All right, let me just make this a little bit bigger. There are popular ones down there. Amazon, Google, Snowflakes is actually quite quite an interesting organization of quite a quite a flexible set of services.
You know something to know what Snowflakes are quite quite quite talked about now. So these are typical the AAS data warehouse as as a service.
So they provide the services whether it's in terms of computational power processing power storage whether it's resource or server management.
So they have your own service there and all you need to do is just use some tools that they have so you can actually have a complete reliance 100% reliance on one of this.
So you're storing your data there. You're using you're not having your own engineers because they're just too damn expensive.
All you're doing is you have really good people that can actually go in and use their tools that they provide for you, for example.

And yeah, or if you go, you can do all the analytics you want. All right, so some key differences between the red shape which is part of Amazon is they tend to be more self-managed.
In other words, you are more self-managed and you will need your own data engineer. So this will be for the big organization to have bigger teams.
But if you are if you're kind of smaller a small SME or a smaller type organization.
This is a sort of technology that you would be going for, which is also known as serverless.
All right, so this is where management is taken care of by the service provider by Google and snowflakes.
And all you need to do is so it depends on the kind of services that you need and you will be spelled up quite clearly in your in your contract.
The pricing therefore is quite challenging because it's based on many different variables, different factors such as the excess that you're going to have per hour.
So how is and your charge your by the hour, the number of concurrent users that you're going to have. So you may have like, you know, 20 uses, 100 uses concurrent means using the same at the same time, trying to integrate information from the database, could be the size of the data and so on and so forth.
So this is the way that the world is heading towards is ready to us and more of a cloud technology simply because we don't have the skills now, especially in the world of data analysts and engineers and so on.
And also because I think it's quite competitive in terms of the pricing, which is good for consumers.
But that's as much as I want to say on on that and hopefully that's that will be new information.
Let's see just checking on my little clock here.
Okay.
We're doing well.
Let's go on to start to look at what 402 is is really about although it's very important that we took some time to cover some of the stuff that we we covered because I'm not going to go through that again.
Hopefully that will start to make sense now because hopefully I've managed to consolidate all this
But really I want to look at these things here from data cleaning transformation integration reduction and
These are all different things that these are these are things that data analysts must do or they have to do is really part of their their role.
As I said, when I started in this business, I did a bit of this, but in those days that weren't these terms, okay, it was all about is the data ready.
That's it.
And you know, you're one and two men show and it's just quite difficult to sort of look at the best practices, especially when there wasn't a lot of analysts around the place.
So now we're in this position, we're now in 2023, this this this whole field of data science is really expanding.
And therefore this becomes really important now is we're talking about.
Okay, so I'm going to talk about one after the other and hopefully I'll show some examples and make sure that I won't fumble in this because I'll be running around different looking at different.
Excel and and and and and and and spider and so on and looking at slides as well.
So let's look at this one here, all right, next one.
All right, make that a little bit bigger.
Oops, too big.
Okay, so at this stage, let me just just to instill this in your mind, what structured data is or stuff.
It's really about highly organized information that can be included in the database usually goes in columns and easily search with simple search operations.
Where there's unstructured data is essentially the opposite.
Okay, really hard to find examples of unstructured data would be your email, for example, very different structure for for everyone.
So emails are very unstructured data.
So the problem with with unstructured data is in the world of big data is it's huge volume.
It's just huge volume.
Okay, but this is where data science becomes interesting, but this is where data science is useful.

All right, so we look at some of these techniques here.
Now let's start off with some of these information about why why cleaning is required.
Why what mix data dirty is another way of phrasing this particular question here.
Well, in most cases, the data that I've been exposed to most in a lot of them can be incomplete.
So really we have attribute value. So we have ratings labels that are lacking.
We have certain information that the manager wants, but it's just not there.
So somehow it's not maybe it's not captured.
It's was was left out. Okay, so that is pretty much incomplete.
So we need to clean this based on requirements and the reasons why we're doing this.
Sometimes it can be noisy. All right, noisy in the sense of of data science.
It's really about having data that contains errors.
All right, errors or outliers. Now outliers are sort of information or data that is just ridiculous.
Let's put it that way.
So if you're measuring going back to my background, which is held.
If I was measuring and collecting data on people's temperature.
All right, so you would expect people's temperature to be around 36.X, 36.1, 0.2, 0.3, you know, 37, 37.1.
But when you see something like 301, it should ring a bell.
You know, temperature of 301 means you're dead. I mean, that's an outlier.
So somehow they could be an error and operators error or an input error.
All right, so that's that's what I meant by noisy. That's noisy data, which are errors and outliers.
Data can also be inconsistent in terms of, for example, how you actually name variables, for example, in
All right, so inconsistency in in courts or names.
So for example, I'm just reading what we have here. The name column for registration records of employees for staff, for pains values other than alphabetical letters.
Okay, make sense.
In organization, some of them will will part of the policy is the last name should be should be in big letters.
Okay.
In some organizations, they have different stipulations. But if you see a value under name, it's very clear that there is a bit of inconsistency going on in fact, it's an error.
All right, so this is some of those reasons that this becomes very important.
Now, there are a couple of ways in which we can we can start to clean up this dirty data data with all these different noises and so on in consistency.
When I talk about clean, I'm one I'm saying is better organize to scrub off the incorrect information or data data that's incomplete, we make it complete or try our best to make it complete duplication data for whatever reason we need to get rid of those duplicated data.
So when and so forth.
So the first thing we need to do is something called data mugging or wrangling, all right, these are terms that you would you would actually see in in the literature.
But basically what it means is that the data is not in the format that is easy to work with.
All right, so you want to make sure that data is easier to be understood to be easier to be analyzed.
So consider this data here and this is data. This is what we call unstructured data.
So when writing down saying add two dice of tomatoes, three cloves of garlic and a pinch of salt in the mix.
Well, that sounds pretty logical to us. Most of us will understand exactly what that means, especially when you love to cook, right, but that is still what we call unstructured, all right, because you could write it in
After wrangling or mugging, it should look like a little bit like the bottom.
Okay, so what you're trying to turn the data into something that is more more convenient to be able to understand and to be able to read.
So in other words, what we want to do is to turn this into a table, all right, so the table conveys the same information as the text, except it is more analysis friendly, isn't it?

Now we have three columns, we have the ingredients, we have the quantity that we need, and we have the sort of dimensions or the size.
It's easier to read.
Okay, the question then is how did that sentence get turned into a table?
Common sense, that's what it is. There's no magic answer to that, all right, it is just common sense, that's it, all right, so that's one really easy way to do it.
Another way is to make sure that we want to handle missing data.
Now, missing data is so common, all right, this is where sometimes we can replace missing values with NA, not a applicable or sometimes we just leave it blank or sometimes we just put a zero.
Now, who knows what that zero means, but to the person that put it in, it makes sense, all right, but to someone that tries to understand it, zero, zero could be an actual value, all right, so these are the things that we need to really consider in this sense, all right.
So the question at the bottom, a lot of these are due to human errors while they're trying to store or they're trying to transfer the data, right.
So what to do when encountering missing data, what do you do?
Well, there's no really good, one good single answer to that, really.
We need to find a different strategies and again, going back to your organization, they have your own strategies, I remember when I was working for different groups, we have our own strategies that we have discussed among ourselves and we have agreed upon them to try to to to to to combat this missing data, which includes things like I'm going to ignore that record, all right, it is missing, I'm just going to delete the
Or I can find some kind of a number or average to actually fill in those missing values, so I can actually change those missing values to perhaps an average of that whole column and that's a very common way to do it to in terms of handling missing information.
Okay.
So what do we have here?
The third method to clean dirty data is obviously to try to smooth this noisy data out.
Now, what does that mean?
Now, there are times when data is not missing, they're not missing, they're not an aid, they're not zero, but they actually, they actually corrupt it.
And this is really hard to to figure out because let me give you an example over here, which hopefully will make make sense.
Before I do that, I just want to just point out to you that a lot of these corruptions are caused by maybe faulty data collection instruments.
I could be by the user, a dentry and a data entry problem in an era in that in that case.
So let me just just talk about this in terms of again, back to my industry, about trying to smooth out these noise again going to taking temperatures.
You know, sometimes when we start to look at temperature, we can, one of them could be an input could be In Celsius, this is a temperature of 37.1 Celsius.
All right.
And now one could be 3037 Celsius.
And then one could be 37.9 Celsius.
Now, when it comes down to 37.9 Celsius, we kind of know that this person is having a fever.
But all right, that's not a normal kind of temperature.
All right.
But sometimes what's happening is we tend to miss represent temperature, for example.
And we start to try to to clean this, for example, trying to change this to maybe one decimal or two decimal or even zero decimal.
In our example, if we were to put this all into one decimal or maybe zero decimal, those values become 37 degrees or Celsius, which means this person doesn't have a fever.

So which means that there is some noise in the era now because that's not true because we know that other the tree values that I just talked about.
One of them was was pointing to a person with with with a fever of 37.9.
Nice Celsius.
Okay.
So that's what I was trying to get across as an example.
Yeah.
I guess the fix down there, the point on the fix is to identify these these outliers.
If you can spot them to remove them if you can.
And the second thing you can do is to resolve in consistency in the data in the data.
All right. So for example, for the inconsistency part.
Let's say all entries of a customer name in the sales data should follow the normal convention.
For example, make sure that all the surnames are in capital.
If that is not, if those names are not in capital, you need to go in and clean that up.
All right. You need to resolve those inconsistencies.
Okay. I think I've enough.
I've said enough about that.
And hopefully that makes sense.
But let's make it more easier by looking at some some examples here.
And you can work along with me over here as I go through this data set here.
Now this data set. Oh, come on.
This data set is a small one, but let's assume this is a huge one.
Where is this?
Do you have it?
Let me just check if you do have it at the stage. Give me a second to have a look at your.
Your learn line site under data sets.
Yep, it's here.
All right. This is the this is the this is so I put it there.
So this is the Excel data set that we're going to play with.
You can pull this up and have it on your screen if you want to or you can play with this.
I'm practice with this a little bit later, but if you click on it, that's what you get.
Let me see. I'll click on it and this is what I get.
There you go.
Okay. That's what it is.
Correct. So in my case, I don't have to worry about that.
I just need to go back to where I was, which is.
Which is where it is here.
All right. So that's what it is.
That's a little bit bigger.
So let me run through this going through some of those concepts that we just sort of heard about.
All right. So that's the that's a data set.
That I have on Excel, but let me try to explain this over here.
So this is about excessive wine consumption and mortality rate.
All right. So on this page here, it talks about some of the labels.
I should have put it down here, but if I were to look at the next slide, it does show me the.
Gives me a further explanation of each one of those labels.
So the first one is the name of the country from which the sample is obtained.
The second one is the alcohol consumption measured in letters of wine per capita.
So when we talk about per capita, we're including in the population.
So to calculate that it is the total amount of the total consumption of wine, for example, divided by the

That's what it's how it's measured.
The next one is the number of people dying from alcohol consumption per 100,000 people.
D is the number of hot disease deaths per 100,000 people.
And the last one is the number of deaths from liver diseases.
Okay. So you can see up here it's liver.
People dying from hot.
People dying from alcohol.
And this is the alcohol consumption. All right.
So this is a typical data sets that's given to you.
This sort of compress. This is, you know, per se.
This is not big data, obviously.
But you know, for the purpose of of practicality and looking at it in an easier format.
This is this is what we'll work with.
Okay. So there's 21 countries here.
If we pull in all the countries in the world, there definitely be more than 200 countries.
All right. So what can we do with this?
We want to find out going through these these these steps that we kind of just talked about in terms of identifying noisy data.
How do we fix it?
Identifying missing data? How do we fix it and do a bit of wrangling, which is probably quite quite basic.
If you need to look at things that you need to be fixed, fix it.
Okay. So let's go through this exercise together.
And hopefully we can we can benefit from this this exercise here.
So back to my Excel, because I can't do anything on on a static site.
So if I go back to my my Excel and this can easily be done by Excel, rather than going through a python.
In fact, a lot of things you can go through through Excel.
Excel is probably the best cleansing tool.
At the moment, the most people will be using in terms of getting rid of the noise, getting rid of outliers trying to find out how to outliers.
It's a lot of it's visual. All right. So a lot of it is visual. Therefore, I want to make it a little bit bigger.
Okay. Let's start with something.
Let me just have a look at this stuff here.
Let's look at trying to look at the noise.
All right. So we're trying to smooth out noisy noisy data now.
And try to make sense of this. Let's make this a little bit bigger.
So let's double click on this.
We'll include all of these information.
We have noticed that in here we have NA.
We have two NA's here, which we try to fix true.
And according to my organizational policy, there shouldn't be any NA's. All right. So I need to fix that.
I also need to go through this and make sure that these numbers are sensible.
All right. So this is alcohol intake.
A little per capita.
So obviously in Australia, it's 2.5 liters per person.
That's what it means.
At the moment, I'm not going to worry about formatting. I'm cool with this.
But I'm just going to go down and see what can be wrong.
Okay. This is where we're going to have participation.
What do you see is wrong or potential?
If I go down the alcohol, for example.

What looks a bit strange?
Anyone?
Yeah. Okay. Good. Good spot.
So this looks a bit strange here.
It does. All right. Because you can't have a negative alcohol consumption.
All right. So my assumption there is this is an error, probably by the person that actually inputs information.
And it's okay to mix some assumptions and it's okay to say, well, guess what?
I'm going to change this. I'm going to get rid of the negative.
All right. And it seems logical.
It seems like there are some 0.8s here. So there's nothing wrong with that.
All right. What else? That's.
I'm just following on what you're saying.
Yes. That seems a bit of a rubbish.
You kind of a negative here. And again, I can get rid of that.
But then as you sell, does that make sense 834? Yeah.
That seems seems like a sensible number.
We have a thousand here. So what, you know, and it looks like it fits in there quite nicely.
Okay.
Missing data. Let's try to work that one out now.
That's more places. Eric, I will clean it up later towards the end.
I'll keep that in mind. You can. Obviously, you can.
Because we need to do some integration towards the end.
I just want to make sure that's.
But for now, let's let's keep the decimal places to what it is, which is a bit, you know, all the way to place.
But thanks for that. Thanks for pointing that out.
Because that was the first thing I looked at.
That's a lot of spaces.
Okay. Now what else we want to do?
What's your take on an A's now?
We talked about that. You can either.
Blanket off. Let's get rid of the colors.
Get rid of that now.
So you'll make more sense.
What are some of the.
Any, any.
You can. Subtract or delete.
Go back to source. All right.
You know, again, this is, you know, it can be.
It can be quite open, guys. You know, there's no.
One correct way to do this.
Some organizations will say, well, I'm going to.
Take a blank.
All right. You can take a blank.
Just get rid of that and that's it.
Because when you calculate, it's going to exclude the blanks anyway.
All right. Forget about putting an an any because according to my policy,
we have to fill in something or leave it blank.
Can I put a zero in there? Is that a good idea?
No, it looks really good. Right.
So I'm not going to put in zero in there.

Now, very common.
We put an average in there.
All right. So what's the average of this?
So I can just put something here.
I can say the average of.
This.
Who's.
Did I spell that right?
It doesn't look.
An average of.
Of these numbers.
184 point something.
All right.
If you round that up, it looks like you need to round that up.
It's about 185. So I could essentially put it in.
I can.
That's one way of being in.
Right. Again, based on your organization and against your policy.
So when you sit down with your team to decide on these sort of things,
make sure that everyone agrees on it.
All right. Let's say we take the average.
Instead of living living a blank or zero, we'll just put the average there.
Obviously, this is rubbish.
I'm not.
This is not true.
Get rid of that one.
And over here, you can do the same.
I can say, well, what's the average of that?
For my liver, people dying from liver diseases.
I can then say.
Okay.
So an average of that is 20 and maybe I just substitute that with an average.
20 point.
Point three.
Okay.
Again, we'll clean up the decimal.
I still remember what.
What Eric has said.
Okay. So that's where you end up with. All right.
Yeah, at this stage, I think it's it's okay.
It's all.
Handle to a certain extent. I think that's that's fine.
That's anything we need to to look at is obviously.
If we start to go down here again, looking at errors, looking at noise again.
Looking at alcohol.
All right. Going through these figures down.
You will now, but let's look at outliers.
I just values that I just.
Totally outstanding. It's just different, right?
They're just out of the.

You know, they're just out of the typical boundaries, the upper and lower boundaries.
Now, I can start to to see it here now that it leads a bit of a problem.
And I'm just going to put this thing here.
Because it's it's 27.9, which is way above what most of these countries are.
And I can also see that this is just no way it's just a bit too small.
I know these are examples and I'm just using this as examples.
So what we do is how do you fix this?
You know, again, you can use what you can use average.
And it's fine. Again, organizational policy allows that.
That's okay. But but looking at this closely.
Looking at this closely and the fact that I just just find an average here.
Just to get myself a better understanding of what these numbers are all about.
Okay, never mind. Oh, yeah, yeah, those are the original numbers. That's that school.
Although those are two outliers, it comes up with some average of probably lower than three, three point.
Oh, okay. So that's that will give me an indication of just a really rough kind of figure.
Yes, Ingrid says very good point. Ingrid about searching the internet for alternative data.
Yes, you can even be looking at maybe last periods data.
Maybe this is done every every year or every month or every quarter.
Look at those figures and you know, make some good sense out of it.
Another possibility of doing this is looking at this itself.
And saying and asking yourself some really logical question.
I'm saying, if this was a.
And input error. All right. So someone must have missed type on the key board or something like that.
And if I take that into consideration and only take that into consideration.
All right, I can put down well 27. You know, maybe the two is is something that.
You know, maybe it's a mistype was typing the bit too quickly, but if it was 7.9, that would be a logical value
Potentially could because France at 9.1, that's pretty high.
But you know, maybe at this stage, the person has.
It's gone a bit, you know, maybe you know, like a little bit of my fat fingers here, typing that.
So maybe I can remove that.
That is something that could be now this is this is taken. This is this is a process where.
You would need to go true. Right now some organizations may not support that sort of thing.
But you you can.
All right, or it could be yes, as Glendice said, it's 2.79.
That's the other way of looking at it. Fantastic.
So using this kind of method may not be the best way to do it.
So it's 2.79 or in my case, 7.9 is a world of difference, isn't it?
So you've got to make the call. You know, the longer you work in this business, the better you get.
And you get a whole because you're going to understand this data again.
And again, you've seen it so many times and you know.
So I'll leave that to you to to so I'm going to take Glendice recommendation here and say to 2.
Again, there's no correct or right answer to this.
But just a basic understanding of why you have put it there. Right.
How about this one here?
What what do you think the possible error could be keeping in mind that you know, Iceland, Iceland, Norway and Finland are pretty much in the same location, which implies that these guys don't drink too much.
You know, the values are pretty much the same.
So given that the fact that Iceland is 0.8 and Finland is 0.8, I would assume that Norway would be a 0.8.
So at this stage, this would probably be an error of one too many zeros.
I can go back and get rid of that. And that would be my value.

And you have to be very comfortable with that. All right.
And you have to live with this as an analyst.
All right. And hopefully you document all these things down.
All right. So again, back to your domain, back to your knowledge, your knowledge base, which is probably sitting in your head.
All right. So that's getting rid of those what I call out liars.
Now another way which we look at into us, the end of the units in the end of these 12 weeks is plotting this.
And you graph this. These things will stand out like a short time isn't it?
It's right at the end.
And it will give you a very good indication that these are outliers and perhaps what do you do with them?
Okay. What else do we have that we can talk about in terms of what the goal here is trying to make sure that this is nice and clean.
I'm just going through it now.
All right. And if we're happy with it, let's let's stick to that.
Let's let's challenge you a little bit now.
You know, you have done this thing. You've done it to a certain extent. You're quite happy with it.
You're looking at numbers and you're saying, yeah, you have this is a bit small. France 11.
But you know, I think that's that's okay because the fact that you know.
Because there is some kind of relationship between the drinking of fabric of alcohol that relationship between that and the number of people that is dying.
I don't know. But you know, it seems again based on on previous data sets.
You can look at this figure. Perhaps.
Now, let me just throw you a challenge. Let's say now you're quite happy with these thing. You're going to do your analysis. Your boss comes to you and say, hang on Lawrence.
We have to put in one more country. All right. We need to put in India.
All right. India has just come in and say, this is our data. Sorry for the delay. But please add us in.
Obligation. Yes, I have to do it. But I don't have the data. But I have the data from India.
All right. So this is what the Indians have sent.
So the Indians have sent you something that's totally different from what you have been playing with.
These are obviously seven Indian states, some of which I can't even pronounce. But there are seven Indian
There is an alcohol consumption. So I'm looking at the labels. I'm trying to make sense out of this.
Okay. I'm looking at a heart disease and say this are these are numbers. They should are very different from
My numbers as in these numbers. You see, there are differences.
And also notice that they don't have numbers on the liver. So what happened to the liver?
These numbers are a little bit huge compared to what I am used to be playing with.
It's a challenge, right? So what we need to do now is to do a few things. We need to incorporate this into
And fill in the blanks. The best of our ability as an analyst.
Now, the good news is India has given us some information. So I need to go back to that information, which
No, not that one. This one. All right.
So this Indian story that we're looking at is where we are given another data source, which is also about alcohol consumption and the number related for fatalities across seven states of India.
That's what we're given. But we need to integrate. And that's where this. This thing comes in. We need to integrate that data, which is here into our integrate into our original data set.
By purpose or by intent or by objective. We're also given information below name of the state, letters of alcohol consume per capita. So that sounds like that is a sort of that sounds like it's quite similar to our label,
The number of fatal heart diseases measured per million people. So that's where the difference comes in.
In India, obviously, they got a huge massive population. And that's why they intended to measure it in a million. The number of fatal accidents related to alcohol, not too hard, not to deliver. Also by the millions.
So given this sort of information is sort of says, okay, now I can't understand why these these numbers are a little bit back. We need to sort of reduce them down to what we have.

Alright, so we're not going through the actual working truth, but just a sort of a just a just a talk walk through. So given that information here.
I'm going to do an average here just to find out.
Because now we're looking at one India, not six or seven states.
I need to give myself a nice picture of exactly what's happening here in India.
Alright, so the alcohol consumption across seven and I'm making an assumption here that I'm just going to average them. So a 2.52.
Alright, and that sounds like a really good old and go a little bit high at 5.79. But 5 2.52 seems and I can go here and say, what for God, the number already 2.52.
Okay.
So I'm going to give you an information for the liver. Let's put a question mark here. What are you going to do? Obviously, it's not a good practice given the context of this data set.
Perhaps it does work. Perhaps it does not. Again, what do you want to do here? Anyone?
Put some value there which we have to put a number. We cannot keep it missing. We cannot put an end there. And a we cannot leave it a blank.
What do you suggest we put in there for India?
Send it back average, average, can we tell them? Yes, we can. Yes, I mean, that is logical going back to them and say, hey, you left out your your your your liver's information.
Can you get it to me? And they say, oh yeah, that will take us another two months. Oh, geez, I don't have
You know, research your previous data set. Yes, you could take an average. You can none of those are.
Are wrong. That's what I'm trying to say. But again, I'm just trying to tickle your brains now to make sure that you kind of think about this.
And then this two, which I'm not going to calculate again, given this original data, it was all per what?
Per 100,000. All right, but India's data comes in as per million. So you need to convert that over. So again, if you were to go to India, what you probably do is find an average here.
All right, you find an average here.
It's going to do this, but I'm not going to do interpretation of those numbers.
Then you're going to do an average here as well.
Can drag that across. Okay. All right. For me to read this better and it's a good practice again, you need to
Calmness is easier for analysts. Get rid of the zeros. I don't want that.
All right. That would be in that situation. Okay. So given that information, you will then go back to your.
Very much. You then go back to your original.
And you can just fix that in.
Yeah. Okay. That was what I wanted to go through guys. And hopefully it's helpful in terms of exactly what was trying to trying to say here.
Let me see. I'm just going through my notes here.
If I need to look at anything else. Oh, yes. The other thing which Eric was trying to say is maybe we need to do a bit of cleaning up here.
Let's get rid of that stuff. Whoops. What did I do?
Okay. Let's get rid of those.
All right. Assuming all this will all fill then.
A bit of cleaning up. Perhaps it would be good at this stage.
I would want those not to be a general. Maybe I'll turn those into a number.
A number. Let me expand that. So I want that to be maybe in three decimal places. For example.
Yeah. Something like that.
Okay. And you can do a bit of cleaning up here as well.
Change that to a number.
Maybe that one to decimal would be would be good.
Okay. You can kind of do some cleaning up. Make sure it's all nice and clean and looks presentable before

You're going to submit this for analysis. So what we're doing now is what the steps that we kind of went through was really about pre validation is about is about pre.
What was the way I was trying to create pre analysis stage. All right.
To make sure that you get all this up and running. Make sure that you get rid of all the noise.
Okay. Good.
The last thing I want to talk about and back to our slides here.
Is some kind of looked at this.
Again, some some information here.
So that's what I meant by that data transformation. I'm transforming it.
Okay. I'm transforming it to another variable. Just like what we transform in the India situation where we have per million.
We need to convert that to a per 100,000. That's what I mean by transformation transformation also includes things like between temperatures, for example, from Celsius to to Fahrenheit.
All right.
Data reduction is what I is here data reduction is a is about what I've just done is my last task, which was reducing the decimal points.
So I'm going to do a form of data reduction in that a form of data reduction is as you can see here in
We have a I guess he could be time, whatever we have 200 of those 200 rolls here and we have 200
And we're trying to reduce it this down to 150 rows and 120 columns. How do you do that? All right.
So we have a constant criteria common knowledge through instructions, whatever it is that is that that reduction is a form of of in terms of columns and rows.
Now the last thing when I talk about is before we go to Python and play around with a data data set for your assessment is discrete comes with what discrete discretization.
So I'm going to give you an example here, which I shoot right now, maybe a ran out of time or space.
But basically what it is is let me give you an example. Let's say we go back to our alcohol example here.
And I want to look at the wine consumption column.
And the boss comes back to me and said, now I'm not I don't really care about the values.
What I'm going to talk about is, you know, if you can sort of categorize that for me.
So he wants a categorization of a variable.
All right. So in terms of temperature, you can categorize that into maybe low and high according to your, you know, whether you're fever or not.
In terms of businesses in terms of sales, you might you can categorize your values into low, medium and
And in terms of alcohol, we might, you know, just think about it as an example.
You want to just discretize that into the boss says for categories.
So I assume I can represent the categories by zero, one, two, and three or one, two, three or four.
And I will say things like if it's less than, if it's less than one.
I will categorize them as as one.
If these values here are, let's say more than one, but less than two, I can categorize them as one.
If these values here fall between two and five, for example, I can categorize them as.
As tree, correct. So on and so forth.
I hope you get the idea, guys. This is where we start to put them into ranges.
All right, like zeros, you know, just like.
Example another cooking, give me a good example where you can categorize things into, you know, into maybe low, medium, high or zero, one, two.
Anyone, but they can categorize any variables.
They come up come that you have actually come across.
And that would be a good example of data discretization.
It means to be discrete to be able to categorize them.
Any good examples there?
Hopefully still around.

Anyone?
I'll let you think about it. Okay.
Gender? Well.
Well, yes and no, gender is, yeah, I guess you can categorize.
If your data is showing me zero and once.
Right. I guess you can, you can categorize into different genders given the kind of genders differences of the gender forms that we have nowadays.
Yes and no to that. But more, yes, they know.
You can be categorized. Yep.
Much another one.
Oh, yes, yes.
Great. Great. Great. Great example.
You know, university. What happens is, you know, you pass, you fail, what do you get?
You got a credit. You get an up the credit is a distinction.
And then you got a high distinction. All right.
Now, whether you get into anyone of those categories depend on your, on your, on your grade.
So, for example, if you have a grade of 85, you know that you're going to be a high distinction student.
Body master BMI. That's another good one. And what about your categorizing to BMI's? Because BMI's a figure isn't an angry.
How, what do you categorize that into?
I have no idea. Poor, poor, good bad.
You know, weight could be something.
Even your height, I'm not, I'm not trying to be getting into all these arguments now of obesity.
It could be low medium high, depending on your, the way that you categorize them weights, you know, you could be, you know, light medium heavy, whatever it is.
I think you get the idea. All right. I think the idea is there.
So, the many times that you will need to do that, I know I've done that many times in terms of trying to be, to try to have categories.
Now, different tools I have to say have different means, different commands to be able to help you to do
I'm very familiar with SPSS. I'm not sure if you know this particular software, SPSS.
I'm just going to pack that down. Another great tool to use SPS. Another great software to know and to learn because I'm seeing this more and more in organizations.
So, believe it or not, SPSS is actually from the social science.
All right. It's a really good tool. In SPSS, I used the word record.
As part of your instructions, you say record. Record, you know, if you are between this range, I'm going to categorize you as low.
If you're between this range, it's medium. That's what it is. Yeah.
It is a statistical tool, a really good one to pick up because I'm seeing that more and more within
All right. What do we want to do for the next couple of minutes? Let's quickly go through this. Let's look at an example in Python. How about that? All right.
So, does anyone have any questions on what we just covered because the next slide is this stuff here, which is really about these are the fundamental slides which came with this particular units.
This is where I asked you to look at it on your own because it has a lot of stuff in it. Try to reach through it and understand it and look at the modules.
I'm going to leave it at that point. Now, let's look at the data. Again, I'm going to I provided you some extra information here.
If you can see it on your screen under data sets, that's one called DVD 40281. So this is what I'm going to play around with now.
And this is the one that you're going to be playing around with when you start to work on your assessment task number one for the second unit.

All right. So for now, if I click on it, it looks like this. This is the real data set. All right. This is what I call big data. And we're going to play around with this right on top other labels.
And you can, as you can imagine, there's quite a number of labels or what we call columns. All right. And these are the actual numbers. The commas here indicates what will give you an indication of what these This is a CSV file. So the values are separated. They are blanks. Yep. They are actually missing values. Correct. They are missing values. They are. Okay. So that's the data that we're going to play with and it's going to close that. Just just keep in mind.
If you had this, just just cut and paste this because that's a useful for putting that into Python. All right. So let's, if you want to do that, you can do that. Cut and paste.
Now we are just going to go straight into Python. And we're going to do a bit of, I guess, trying to explore what data is all about and trying to do checkings. I'm going to go through some filtering some bit of cleaning, for example, and stuff like that. And then we'll finish off for this evening.
All right. So, so hang on guys. We got about 15 minutes to go or maybe even less now. So what I've done. At this stage, I must learn how to expand this on, but I on on the screen. But I think you can see it. Now what I've done at this stage is to.
I was actually going to talk a lot about a lot, a lot more about pandas, but let's leave that for another day. But we have this statement, which you have seen this before, right. Panda is a library within Python, which deals with data manipulation. All right. In other words, it's exactly what we're trying to do here.
Roads and columns, because that's the nature of the beast. That's the nature of our data. So we're going to input pandas as a PD, which is just just a court, court name.
So what I've, what I've done here and we have done this before is to give it a variable name. So in this case, I've given it DF. I'll explain what DF is because I'm trying to explain what data frame is.
So there are new things that I want to bring on board now in terms of our discussion on Python DF stands for data frame DATF R am e data frame is just a, it's just another way of saying, yeah, we have data that has roles and columns. That's what it is.
Now equals to PD PD is from here. So it's telling you that yes, please read for please associate this with pandas, which is a data frame.
I'm a new provider and we're reading this CSV file and this is that file that we were just looking at just now. So remember the brackets, remember the single codes and that's where we cut and paste.
And earlier on, I did this and therefore, if you look at the console, if you look at my Python console, I ran this enter no error. So I'm happy with that because I didn't want to waste time marking around with errors.
I typed this in.
Press this particular symbol here means is to run my individual file, sorry to run my individual line.
And that appears like so, which means that I do not have any error. So I'm quite happy with that. So what I've done is I've actually pulled that data set into into Python.
Now they kind of understand it. All right, let me just explain with some of these new new instructions here as we go along. You can follow along. I can try to understand type.
I want to know the type of this.
DF and the score role that that's the name of my.
That's what I have up here. I could have I could have used an easier name. I could have just used the F, which I should I this is a bit long, but I can use the F over here.
On the left hand side, in that case, I would then type in type the F. So let's run this and see what we get. All right, so we run that and we get an output output is saying that it is a data frame.
All right. And if you look up here.
On the variable explorer and we're going to use this more and more and I managed to I can move this But I need a little bit more space here. So this DF role, which is up here, is a data frame is a type of a data frame rather than an integer or whatever is a data frame, which it has roles in columns.
How many roles we have? So you've asked you a question. How many roles are there in the original data set? That's the answer 50,598 rows. How many variables do I have 79 variables? Okay, that's a lot of variable.

But it's very typical of a data says that that's been sent to you, especially this sort of data set that's very much focused around environmental issues and you know this has to do with carbon, carbon dioxide or carbon emission and so on and so forth.
Now, if I were to double click this and this is just something for you to not, if I were to double click on my name here, it will start to expand this into the actual data frame itself, which is the data set. And as we move, this is the, if I move from left to right here, it shows me the what all the variables, all the columns, all
And guess what? For the first couple of variables, we have nans, any n. Nans is a way of for Python to say, guess what, Lawrence, these are missing values. All right, nans.
Okay, and of course, if you try to scroll down, it's going to scroll down way down. So I'm not going to do that
All right, so that's a nice thing to note. The variable names are also given here. And these are first few.
Okay, what else do we need to to play around with? Well, we can kind of do this sort of things here. We can now and we can do this.
This is Df raw, some of this we have played before shape.
And you can practice on your own. Again, I'm going to run this. And it does give me the shape, which is the number of roles and number of columns.
I'm happy with that. I can also play around with raw columns, which in this case, I'm asking for what other
I'm going to run this line.
And bang, it gives me this 79. Okay, now as an analyst, and if you know this data set very well, it will start to make sense to you.
But for us now, those are also looking at this for the first time you'll be scratching your head. But it's not a gift your indication.
If these labels are written well, and that's why another point is a really good practice for for for analysts and people like you and me to be to give it really good sensible labels. Okay, so that we can look at it and say, yep, we know exactly what this variable is about.
Okay, so that's a that's a that's practicing. So it gives me that that number of variables.
So what we want to do now is we want to I'm going to put a comment here. I want to subset data. So I want to start to break it up into little bits and pieces.
I want to now do a I want to subset the I want to but I want to keep certain certain columns or certain now another word for columns would be certain ratings.
I think that should be easy to understand. All right, so there's a I'm going to use the word keep. I'm going to associate the word keep. So I want to keep.
What I want to do is I want to keep certain variables and I've selected one, two, three, four, five, six, seven, eight variables. All right.
So I want to keep the following variables for my analysis. So this is part of my cleaning up. I don't care about the others. I just want to have these things and what I'm looking at is in single coats. I want the ISO code.
I should have packed it up earlier, but it's okay.
I need the country.
I had to be careful about this. I don't make any arrows because the chances of making arrows is quite large
Arrows especially my fat fingers. Yes.
Population.
And whatever I can't bring here has to be the same as what I see here in the Python console. So this is as population. That's what I need to put POP LATI.
I cannot put in POP, which seems sensible, but you're going to get an error. You do that. All right. You will get
And DPS is the other one. I also won hang on to the theme one to go to and we'll then see all to come in my
Thank you.
And what else is methane?
I always spelled methane.
And I want to keep in nitrous.
I can't go with you.
Oxide.

Okay.

12345678

I just want to make sure these are all spelled the same way as here methane.
Not me. I'm on a nitrous oxide.
Okay. It looks good.
If anyone sees any arrows, let me know.
Okay. So I want to keep these variables.
And I run them and hopefully think across.
Fantastic. Okay. So what has done is the data set is actually cut down my data set into eight.
What do you call it? Eight columns.
All right.
I can do something like this. I can I can do a DF. It's social. DF with.
DF raw.
My DF raw would have changed now.
And I want to look at what I have kept.
All right. Hopefully this will show me some information about what I have created a sub that subset.
So I run this up here.
Okay. It's quite happy with that. It accepted that.
And I want to do a.
Let's see what happens if I do a DF dot columns at this stage.
I'm just going to mess it up. Hopefully not.
Okay. It's. Okay. Just. It's just saying that it's confirming the fact that in my data set now.
I have only eight of the variables and my daughter is there. So I said now.
So my data set now is.
If I were to do a.
Just trying something out. DF.
I'm going to show that.
Yeah. So now it sort of confirms the fact that in my particular my current.
Set I'm looking on now has.
Has that number of rows and and that number of of of variables.
Okay.
And up here in the variable for explore it.
It shows you exactly what's going on.
So keep is actually not a data frame.
Keep is actually a list.
So again, terminology in Python. A list is this is a list.
A list of eight variables and it sort of tells you what the names are.
Okay. So this is what you have.
Keep in mind that you'll be doing some of this when you go through the assessment because I'll be asking
To pull out certain variables and stuff like that to look at what you have and to confirm.
Now let's say we want to do one something like that before we finish.
I want to do more subset.
I want to have a.
I want to exclude.
Exclude 18th century data.
Because I know in there they were they were.
That's our way back in way back in time.
All right.
And I don't want to have those things.
And in fact, I want to.

Yeah.
So another way of saying.
If you're including 18th century data.
What you're saying is I want to include.
Data that's from 1900 onwards.
Okay.
So how do you do that?
So there's a statement here that you can put.
The F will be now the F some again,
again, sub setting that.
The F.
Following the notation.
The F dot year.
All right.
My year is more than or equal to 1900.
Post that with a square bracket.
Okay.
And again, these are these are Python.
Conventions.
You need to stick with a syntax.
So this is what I'm saying here is.
You know, I just want to keep in my data in my data frame.
All the years and the year is here.
As you can see here is is a variable here.
Making sure that my years are all my data will only contain 1900 and more.
I don't really care about any information that's less than in a 1900s.
Okay.
So you were to do that.
And do a run.
That has no errors.
And now we can actually do a DF.
God.
Shake to kind of have a look at it.
See how many records that were gotten rid of.
All right.
So from 50,000 now is dropped down to 33,000.
So okay, just keep that in mind.
So what we're doing now exactly what we were looking at in terms of conceptually.
And the last thing I want to say.
How do you chat to make sure that this is correct.
How do you actually check.
Check.
Yeah.
How do you check that your information is is all 1900 onwards.
You can say you can say what's the minimum of the F.
What year.
What's the minimum year in my data frame now.
And if I press on that, what do you think that answer would be.
Anyone.
If I run that line.

Yes.
If I was if that was correct.
If I do a run here.
Right.
It should say exactly what Glenn is angry at a set 1900.
Okay.
And that gives you a sort of a comfortable feeling that you are actually.
So what you have started here as an example, you started off with a massive data set.
You're given some instructions by someone or some organizations.
Or it's a purpose or intent that you need to reduce this down.
You need to get rid of the noise.
You need to be able to get rid of duplications.
There's no duplications here in this particular data set.
But there was a bit of noise.
Like you know, 79 variables is a bit noisy.
So we kind of reduce it.
We're going to talk to the people.
What kind of what kind of analysis do you want and then we work backwards.
We'll then collect the relevant information like so.
And we start to make it.
You know, make it cleaner and cleaner and cleaner.
Alright.
So that's where I want to stop now.
And hopefully the time that we have just checking my time.
That's an excellent time.
So that's this where we are now.
And then from here you can save this this.
Very clean data set.
Save it.
Onto your working directory.
Give it another name so that you can come back to it and play with it.
Or present it for analysis.
Okay.
So what I've done today is actually a lot on validation.
Isn't it guys?
A lot of making sure that you get your data prepared for.
For pre analysis, which is unit 403.
Where we actually use the stats statistics and the modeling to do certain things that again would help us to achieve certain outcomes.
Alright, at this stage, do I have any questions from any one of you?
I know it's been long.
I know I covered a fair bit of things that hopefully that was helpful.
Because that that's as I said, it's already best practice that we're looking at now.
No questions.
All good.
Yep.
Alright, so based on that, I assume that you all pre-tired.
I'm tired.
So let's call it quits.
You just spend two hours together, which is great.

What we'll do next week is we'll look at the first 81 of this second unit.
And we'll move on from there.
Okay, so again, thank you very much for your attention.
And I'll look forward to talking to you again next week.
Very much.
Just guys.

0

Lecture 6 - XBD401 T3 20230830 1632 Tutorial 1

All right, so good afternoon. Good afternoon, Lisa. Hope you're doing well.
Yes, all as well. Life's very busy, but we're fitting everything in.
Yes, for some reason, life is just busy, isn't it? Just busy, busy, busy.
All right, so today is the day that we're going to look at the first assessment of XBD402, which is really about testing big data.
And what we did last week was to go through some of the concepts, and I went through a series of exercises using BOT, I guess, Excel, and also Python to show how some of these activities, like cleansing and transformation and so on, can be performed. And there's a typical sort of task that an analyst would perform almost on a daily basis.
All right, so what we wanted to do today, let me just switch over to my actual questions here.
All right, but before I start, just just a reminder that the due date for this particular unit here, 402, the date is the 12th of September. All right, so that's just a due date. That's just a date that is written in your study plan. If you submit this 82 and 81 by then, that will be great.
If not, just keep on going at it, because as I said, probably a number of times now that towards the end, that tend to be a bit of a stressful time.
So just be consistent in that respect. Just a bit of marketing, I guess from my part, before I forget is this skill set will not be offered in term four of this year, because it's just too short. But if you have friends who are interested in knowing a little bit more about big data or data science, we are going to offer this skill set every term for next year, just the first three tabs. Term one, two, and three. All right, because we are seeing a large number of demand, we have students that are quite interested in knowing a little bit more about this particular discipline. I guess the other area which I like to just say is also cyber security management. So that's also been offered in term one, two, and three of next year. It's also very similar to big data, except the focus is on cyber security management.
So in itself, it's like a bit of a skill set, very similar structure, 12 weeks, four units.
And that is also, I believe, what are these skill sets are subsidized by the government, which is great. All right, so if you have friends who are interested in cyber security management and or slash all big data, yeah, let them know and we'll go from there. All right, before I start looking at the assessment, I'm going to ask Lisa whether you have any questions on what we kind of covered last week. And I believe you were there.
Yes, I was there. No questions as yet. Okay, great. All right. So I'm going to move on from on looking at this first assessment. What I've done is I've just made some modification to it quite recently. So if we had downloaded this version, the original version maybe two weeks ago, get rid of it. This is the latest version. The latest version has, I kind of reduced the number of questions from seven to six. All right, so question six is a sort of a hands-on, is a python question. But the first five are more like concepts. All right, so some of these concepts are concepts which we may not have covered last week, but that's that's fine. It's just a learning process. And what I want to try today, something something new is, well, we got Thomas trying to join us. I'm going to actually use chat GPT to to look at some of the the questions that we have. Hi, Thomas. How are you? Yeah, good, Mark. There you go. Okay, I'm good. I'm good.
So we just started. So no worries there. What was I saying? Oh, yeah, chat, chat GPT.
So I'm going to use chat GPT and run through some of these questions through chat GPT because I I find that chat GPT is a really good learning resource. That doesn't mean that I'm trying to promote a different kind of learning where you go in and actually zoom into your questions for your assessments, but it's more like going through this chat GPT this time around, I will identify some of the things that you should not be reading in there and stuff like that. So you'll be quite interesting as to to see how this this afternoon session goes. Okay, right. Let's look at the first question here, which is very straightforward. Number one and 1.1.2 is around the data protection law related to testing big data and also privacy law related to testing big data. Now,

per se there is no specific legislation just on testing alone. It's pretty much around what the what the laws are in terms of the privacy act. So we always go back to the privacy act, which tends to the privacy act of 1988, which regulates these entities, which I talked about weeks ago, what I call the APP entities. All right, APP stands for the Australian privacy principles if you can remember. So some of these APPS most relevant to the testing function of the big data. All right, so just to put this into perspective, what is testing? We kind of talked about it last week, but testing is really about in the context of big data is really about making sure to ensure that the data is accurate, is reliable and it performs. All right, before we actually do the data analysis. So again, ready the data for analysis. So before that, it needs to be clean, it needs to be accurate, it needs to be reliable. Those sort of things is what the focus is on. All right, so the baby under this question, what you need is probably the APPS in front of you. If I can just pull that up, I've got it up somewhere here.
APPS, there you go. So I don't have to waste time looking for it.
Let it go. Oops. Quick preferences. Okay, there you go. So these are the Australian privacy principles. And to be able to answer the question, especially the first one, 1.1, you need to go through this list here of the 13 APPS and identify which one of those have relevance to the testing of big data. All right, the cleaning of it, the transformation, we can show that the data is ready for for analysis. So we can go through a few of this and ask ourselves APP1, for example, that has to do with the open and transparent management of personal information, making sure that it's open and transparent.
Does that have anything? Does that have relevance to big data testing? Is there a yes or no? Anyone? Oh, welcome Eric. Can you see you there? Does APP1 have any relevance or association to testing big data?
Can type it in if you want to? Yep. We think there has relevance.
Yeah, Lisa said not directly and I tend to agree with her. Okay. It may have some relevance, but I don't see I'm not totally convinced that it's kind of linked together, isn't it? How about the second one? APP2, which requires individuals not to identify themselves. They can use a quote for themselves. So instead of saying that I'm lawlessly micro-use LL, all right, it's a quote that I've used. So does that have anything to do with the testing of big data? And the answer is,
yes, it does, right? Because one of the things that is part of this cleansing is to make sure that you know, making sure that if this is a survey, is to make sure that your name is not there and you should be coded. All right, so all the different surveys, research groups, whatever research that you're doing, you do not identify and you in fact, be identify the person, by giving them a quote or just a name or a pseudonym. Okay, so what I would like you to do is obviously to go through one of this quickly, one by one and see which one has relevance. Some of these are quite obvious, for example, IPP6, the use and disclosure of personal information, yes, that has relevance and I'll just put that down. So for 1.1, it is a method of going through the 13 privacy principles and noting which one of them has relevance. All right, let's press read forward. 1.2,
describe privacy law related to testing big data. That's just a probably a one sentence or maybe two sentences description of why it is important, all right, because you know, you need to protect personal information, blah, blah, blah, unaltrised access, modification of disclosure of data and so on. So that's why a privacy is relevant to testing big data. All right, so that is that's question 1. If you do have any questions, just shout and I'll stop and I'll address whatever you have. Let me just look at this. Now we have Ingrid on board. Welcome Ingrid to the to this afternoon session. All right, let's look at 2.2, which is defined 2.1 is defined big data testing. Again, you know, you can define it in your own way in terms of how you understand it. It is, I guess in terms of testing is very different from traditional data testing because due to the fact that your data says a much more larger and of course more complex.

Let's see what I did. See what I'm going to go through is a process and I kind of look at this, especially for question number 2, where I got into my GPT chat, it's not GPT, all right, and I hope that most of you would have would have gone in and played around with this this particular AI tool. I use it a fair bit now to learn and to I mean you could do a lot of things with this, but what I've done is I just saw the sake of this particular exercise I have put in almost the exact kind of questions or you know, so this is the prompt, the prompt, so this is what I'm asking chat GPT, please define the data testing, things about it and it comes back with this this answer. Now my advice to you and it's quite lengthy, my advice to you at this station for the rest of your, I guess if you if you are going to university and all that, it's not to cut and paste because that would be that would be to me that that's not good practice, or what you can do is retry this and try to understand because in the first paragraph for example, it actually defines what it is. So big data testing refers to the process of verifying and validating the quality, the quality, the accuracy and performance of large and complex data set within the context of big data application and systems. Sounds good doesn't it? Okay, so I can use that, I can rewrite this into my into my work into my 81, but if you look at the rest of it, it's irrelevant in trying to answer 2.1. So just don't cut and paste, just look for the relevance sections and parts of the answer and you have to just read through and make sure that it does make sense. Okay, if you look at this, it goes on further to talk about key aspects, what have that means? The key aspects of big data testing, let me just make this a little bit bigger. So we oops, that's a bit too big. Okay, and this this sort of information may be useful towards towards other questions, okay, but for 2.1 that that's what it is, just this bit up here, rephrase of cost. Now the second and third one are very generic questions, there are actually no standardized procedure or the proper way to do it, but these are really broad perspectives, like number 2.2 is what are the industry procedure required for big data testing? Obviously this is my answer to that will be, well, it's based on your organization because every different different organizations have their own policies and procedures about doing certain things. So again, what I've done is I got into chat GPT, so keep in mind, questing 2.2 is what is the industry procedure? And 2.3 is what the industry protocols required to write scripts and queries for testing. Alright, so I'll explain a bit more about the 2.3, but let's have a look at the 2.2 where I went in and I did this earlier, so to save time, I just cut this in, and I asked chat GPT, obviously what are the different types of big data testing? So obviously, notice the prompt, the question that I've asked is slightly different from this question here. Okay. What I guess is, it's just an advantage if you were to go into the chat GPT, the prompt should be simple and really easy understood, so that's why I sort of rephrase it by saying what are the different types of big data testing? And it comes up with what I was really looking for, okay, in a way, but not quite, not quite dead because these are just the different types of big data testing. Oh, I know what I did here. I was just curious because the next question is the, is the actual question that I wanted to to know it. So this is good, this one here is actually question 2.2, isn't it? List down the industry procedure required for big data testing. Isn't it? Isn't that right? So that is industry procedure required for big data testing, okay? So that is what I was aiming to do, but before that, I was just playing around. I was just interested in, what was I interested in? I was just interested in the different types of testing. So I asked chat GPT, so what are the different types of big data testing and it comes up with a whole big list, okay? Really good way to learn because as I went through this, I say, yep, functional testing, yep, understand that one, performance testing, yep, I know that one, and so on and so forth, okay? We'll come back to this. And then for some reason, I was just interested in explaining Facebook API, okay? So this is just different ways of just don't worry too much about this. I'll just interested in what they would say if I had asked that question, but back to what we were looking at originally, this is question 2.2, which is to what are some of these industry procedures? And it sort of spits this out. I have gone through it fairly quickly and they say, yep, that sounds

reasonable, but I'm not going to cut and paste the whole thing, right? So what you may want to do is you might look at this as a bullet point. So first bullet point would be, yes, I'm going to get a requirements and analysis and make sure why am I doing testing for? So some of the goals objectives record associated with testing. That's the first step. Sounds good. The second one is profiling on my data. So knowing about the data itself could be useful when you conduct testing. So you try to understand how big it is, how kind of different types of this, the structure, all sorts of things are also quite useful. All right? Then preparing the data for testing. Do I have any test data in mind? So just think back to what we were doing last week where we sort of got the data together and we did a bit of, I guess, trying to understand what the data was, a bit of profiling using Python commands. And then doing some testing as in some of the exercises that we went to last week, such as transformation, cleaning, what do we do to missing values and those sort of stuff? You can, well, setting up the environment is okay, but this is something that you may, you know, because there's so many of them and this is just a chat, GBT's answer. You don't have to list down all 17 procedures. Obviously, the ones that make sense to you just just put it in. All right? So to me, the first one makes sense, the second one makes sense, the third one is preparation for data. Yes, then I can do the functional testing. All right, we should basically the data transformation cleaning. I didn't, we didn't do any aggregation, but that's fine. So really, you can, and we didn't do all the individual testing. All right, as you go along, we didn't do any of the testing because some of this were irrelevant, such as usability testing, which is about testing the user interface of the tool itself. Not, we didn't do that, okay? So it's a matter of going through this and pulling out the ones that is relevant. Documentation and reporting is somewhat relevant, so that could be another bullet point. Continuous monitoring makes sense feedback improvement makes sense, okay? So quite useful in trying to address this thing here, which is basically trying to list down the industry procedure. I believe at this point, if you don't like the answer at this thing, there's something called regenerate. If I'm not going to do that now because it's going to just change this information that I have, if you feel that this is not quite correct, and sometimes it does go into that space where it gives you a bit of rubbish in there, you could click on this and it regenerate a different answer for you, or perhaps rewrite it in a different way. Okay, so try that, to play around with that. What did I do here? Oh, okay, I think I did another prompt here in asking it. So what are the steps in big data testing? All right, so I'm trying to ask the questions in different ways so that I can understand the actual steps of testing. So in this case, it's given me the same answer. All right, but I noticed there are some things different here, such as data ingestion testing. So we talked about this last week. Injusion is really about the that layer behind behind all those different data sets and your data warehouse. So that's where the staging happens. All right, so you can have a look at that answer, but as you can see, it's pretty much similar to what was displayed before. All right, now the next question was this, all right, this is 2.3. So I just took the whole question and just lumped it here and see what whether this thing would be smart enough to figure out exactly what this question is all about. So this is about the industry protocols, the usual way that the industry will work on these to write queries and scripts for big data testing. Now let me just explain that a little bit if you don't have some IT background. Let's go back to testing systems, right? Not just data alone, but systems. So you have a new new system in the office within your department. It could be a system that does, let's say in just just take an example from the university world where we have a system that is a new student management system, which we will be having by the way, a new brand new student management system. And this system will keep track of student information in our demographics, what are the units or causes that they have taken, what greats they have. So it keeps a little bit about US students, all right? As an instructor, I can go in and look up your background, can look up your what have you taken at CDU, have you graduated and if you have in what? So that's a typical student management system. Now let's say if this system was brand new and it needs to be tested and you're one of the

tester, all right? You're one of the tester. So what you would usually do is you'll be given scripts. Now scripts are predefined steps that a tester would go through. So if the script would say number one, press on submit and what I do is when I'm testing the system as a tester, I would follow the script, all right? It could be in steps one, two, three, four, five and I'll go through that. And if there's anything that I don't quite understand or strange, I will write my feedback on the form itself, all right? So feedback could be online or it could be on just a piece of paper. And that's what this thing is all about when it talks about scripts. That's what I mean. A script is just predetermined sort of steps given to you for testing, all right? So you might say check for missing values, step number one. Look at this column A and check it if there are any missing values. So it gives you a sort of a sequential step by step process for big data testing. All right. So now that you know that it's kind of talk about what are some of these industry protocols. Now that's kind of hard to understand but the way that I will probably look at it is what are some of the things to look out for if you are given scripts and you are a tester for big data testing. So some of the things that is written here and I went through this are quite sensible. You know chat GPT you have is getting better and better I would have to say it's not giving me rubbish but it's telling you yeah you should look at version controls. Yes you should look at you know whether your documentation on your queries and scripts are clear for example the instructions are clear that's on the script itself.

If there are sort of you know coding going on and you have the test coding some programming languages and tools that you're using that has to do with coding standards. So I wouldn't win and cut and paste all so long list here again you know it gives you it just fits out a lot of these. Just pick out the ones that would make sense to you and you think there's a bit of red it's relevant to the question all right I would just look at maybe four or five of this if you can but those of you who may not again disagree with this and say no I don't quite like this because you know I work in an organization they actually have protocols already for scripts for testing so I'm going to just use those from my organization that's fine okay so it doesn't really matter at this stage but what I'm going through today this afternoon is just something that I've never done before which is using GPT to actually go through some of this and hopefully I explain that not everything that you read is is totally correct all right because they tend to spit out a whole bunch of stuff okay so again to understand this question a little bit more I I typed in a question which is which was nice and short and simple I've asked GPT to please explain to me the procedure to obtain a big data I think this is question number three if I'm not mistaken just give me a second doing going back to this

now I think I was trying to look at this question over here three point two so let's have a look at three point two now moving along and I was trying to look at the second question here which is to explain the procedure to assemble and obtain a raw big data all right so I'm going to chat GPT and I didn't like the word assemble so I kind of left it out made it a little bit more simple and say chat GPT explain to me the procedure to obtain raw big data and behold it comes up with these sort of yes and if we have gone through some of these the the classes with me the workshop that we have been having this will make a lot of sense yes we need to identify the data sources where are these information coming from I need to understand my data a little bit more all right I need to find out where the sources are and that's exactly what you need to do and then we extract the data using some tools the ETL tools we need to transform them we need to clean them if we need to reduce them we may ingest them if we have a staging layer we need to look at security we need to look at quality we need to make sure that the documentation is done well version control is just a sort of a mechanism for us to keep track of changes so that's why we have a version control such as if you're a Word document you could be improving on it all the time you could have version one version two version three all right that's what it means by version control making sure that we keep track of changes and naming the file is one good way of tracking those changes okay I think those are the key ones again that's that's that's fine let's have a look at the question itself so let's

look at question number three now so far so far does anyone have any question oh Glenda you just trying to welcome so so Glenda I'm not sure when you came in but what we're trying to do is to you know I'm trying to go and use chat GPT to actually try to relate to some of these questions and the main the main point why I'm showing you this is you know some of my students tend to just cut in place that's really dreadful because I know it's from chat GPT but it's just to be a bit smart in terms of reading the output from chat GPT and and and learning from it and making sure that you you know learning from it and and and and writing out or rewrite the answers in your own words um something that Glenda I came in at 81 questions you okay I'm not that's fine that's good uh we have only five questions tonight so it's not too bad isn't it not too bad at all so question three is about again testing your big data resources but it's looking at focusing more on the policies and procedures around it so it would be great if you again have access to such policies and procedures within your workplace all right that helps a lot but if you don't well let's go through some of these things three point one explain organizational policies and procedures relating to testing big data sources so just imagine if you are in the business of testing big data in your organization you have huge amount of this massive data sets sitting around a place just like you know empty health has a lot of these big data sets just sitting around so what are some of those policies and procedures that that you have all right so uh I wrote some of these down for myself this time around I'm not going to go to chat GPT for that you can if you want to but this to me is quite related to your organization so one of the some of the aspects that you would be looking at is obviously governance so data governance is a big deal in organization because this is it drives the standards for data quality to make sure that the data is accurate is complete and so on so I can assure you if you tomorrow you go back to work or even this afternoon if you were to go into your your information repositories and you search on data governance uh chances are you will have data governance policies in there and it's quite interesting to see what they're saying uh you also may have procedures and policies around how you would claim clean data how to transform data and so on and so forth okay you'll you may have guidelines or procedures to help you to handle missing data for example uh what happened if there are huge ridiculous data in there that doesn't make sense what we call out liars what would you do all right so this is all has to do with with with data governance uh I guess another aspect that you can look at in terms of explaining the policies and procedures is around the concept of privacy and security especially when you have very sensitive data where you need some control around accessibility especially electronic health records information and data on on patients is it's just very confidential correct so those are some examples uh 3.2 we talked about this when we talked about how do we obtain big data so again what we saw in in the chat GPT a 3.3 is looking at the process to perform data cleansing um following in in in ETL so basically what is asking you here and it may not be very clear what are some of these uh is it cleansing yeah cleansing data cleansing so you can you can you can list out some of the processes as part of it such as uh yeah I you know I want to some of the techniques I use is to uh is to clean up uh is to handle missing values for example another one is how do I handle duplications and another one is how do I handle uh inconsistent formats for example in my data and how do I transform some of these these variables so again those are all parts of it and I'm quite happy we just want to just put those down bullet points I will understand what you're trying to say all right uh 3.4 is again it's uh this is a tricky one but it's asking you first of all the procedures involving testing transactional versus non-transactional sources of big data so first of all you obviously have to know what the differences are and I think you do right so transactional and mostly uh like day-to-day transactions in the bank for example money coming in and out those sort of information and non-transactional are more like uh more like master files of data so the kind of procedures involving that again it's uh it's a strange question because they are first of all the procedures are different for both of these types of data uh in terms of the transactional data the focus is very much on making sure that the data is accurate consistent it's reliable it's reliable

uh in terms of the non-transactional data sources of data which usually comes from let's say the social media posting logs and so on uh the focus is very much on on general quality all right so that that's where the difference is again it's a it's a very really funny question in this one here uh I do not expect you to list down all the different procedures or or or testing uh and when this word procedure here what is what is referring to is actually this one here let me just point you to so what does that exactly mean if you were to go back to our our our listing here our chat gpt is actually making references to stuff like um the types of testing okay these these sort of things scalability security testing usability um there was a better list up here I saw somewhere up here yep this this sort of stuff here so more like functional testing okay which is about making sure your your data is valid is complete so this one would definitely apply to for example transactional data okay performance testing would would would be referred to in both transactional and non-transactional because of the fact that it's it's doing testing but based on how large your data is so load testing is is basically uh you know the what you do is you will pump different volumes of data into your into the system to make sure that it doesn't collapse all right so you just load tests uh stress test is when you when you have the worst scenario so you really stress your your system with a huge amount of data and see whether it collapses all right that's what stress testing is all about really under extreme conditions such as at a peak load so you have millions and millions and mill even billions of data uh into the uh into the uh into the uh into the warehouse for example okay uh security testing example would be for both transactional and non-so I would go back to this one here which is how did I get to here I just had a question saying what are the different types of big data testing and I've got this yeah I think this is this is quite relevant going through it yeah at least this this it's not a long list it's a pretty good list here all right so what are the different types of big data testing okay
not a 3.5 is about explain the procedure for storing test results and again this can be very generic uh there are procedures again hopefully if you have an organization that you can pull these things down from in terms of what's your policy around storing test results again this these sort of things can be very very generic all right what I would ask you to do is to just type this in I have I have not tried this just type it into uh um chat um GPT and see what it comes up with all right but the answer will be quite generic in terms of some of the things that will look up for if you want to store all right for example is the storage large enough where is it going to be is in a cloud and so on so those are the some of those things that you need to go through as as procedures all right number four has to move along so this is where we look at data validation all right so again has to do with the cleaning and transformation and making sure those things uh if you if you cannot remember what these things are uh I think I have it somewhere here yeah so these are you know things we talked about this last week it's about cleaning it's about transforming it's about integration it's about reduction this discretization all right so you might have to go through and do a bit of review on this but these are all part and parcel of validation making sure that you have validated data before you actually
going to the analysis stage so discuss some of these uh protocols so some of these will be things like the different types of validation and and again I didn't go through this in in chat GPT you might want to go that yet a place that you can get information on is going back to those uh your landline site and into those different modules look for the modules that talks about validation protocols all right so some of the answers are there uh but looking at 4.1 it's just basically discussing some of these protocols uh what to me that means is I will be checking for whether there's for example um making sure that the data is is valid um I'll look at the range all right so for example if I were to look at uh age for example age is my column and I want to validate that so what it means is I could be looking at range so I could stipulate a range of my of how old people are I can say zero to I don't know 150 all right so that's my those are my boundaries those are my boundaries for validation for for age for example which is around range so it has a lower limit and a higher limit I could also test for let's say you're given data there has a customer

ID for example or a staff ID or you know where you need to be unique so that's what we call unique validation so checking for for for feels that has to be unique for example uh there's also validation around formatting so for example the formats formatting means for example take take an Australian home phone number uh typical format would be your your postal code followed by eight digits all right anything else would be classified as invalid or is not validated so that is again some kind of validation all right so just just some examples of of the discussion that goes around validation protocols uh 4.2 again I can give the answer this is pretty straightforward next time I'm going to get rid of this thing it's just a repeat listen explain big data testing methodologies in fact what is asking for is just all the different types of testing all right so again going back to your what we were looking at this is what they're asking for what are the different sorts of of the testing so functional again bullet points functional to form and security you know integration so on and so forth so that would fit into 4.2 you don't really have to explain just just just just list them because I think you have seen that in different questions this evening test script we kind of talked about that is just a set of instructions used to as a guideline for you as a tester all right sometimes it helps us to automate the testing process but basically in most cases you know it could be just a piece of paper with all the instructions on it so that's what a test script is again I know there's some documentation in learn line around in your modules around test test scripts or if you want to google what test script is or going to check GPT or some other AI tools to to find out exactly what test script is but I think you have the idea okay let's move to 5 at this stage let me stop here and ask the group if anyone have any questions so far or or this is going well how's it going guys okay all right good good if not just scream and shout I'll listen the last but but not the least is question five again it's around big data sources all right so 5.1 funny question explain features and formats of common big data from sources what I'm actually looking for here is is the features which are related to structured versus unstructured data that's all that's all I'm asking so you can you can list them and just just explain a little bit about what what what structured data is all about and what unstructured data is all about that's all I'm asking okay at this stage 5.2 again right about big data testing methodology so these are some of these again I'm going to take away this question next next time it's is referring to those those sort of different forms of testing okay which I've showed you this stuff here the different types of testings so that sounds like a repeat of one of the questions and you're on that's okay just just do what you can just fill it in 5.3 are the different strategy behind testing big data so this is where we start to learn a little bit about testing in terms of in terms of time you know so in terms of how it's actually tested so batch testing is I guess the concept of batch is let's say again giving you a bank example I think many years ago what and maybe it's still is it's in the way that it's been done now in banks what happens is your your credit card transactions for example a batch together let's say you have a number of transactions in the day and what happens is it's all these transactions from all the different customers of the bank they are they are run at the end of the end of the day so at the end of the day maybe around 11 p.m. it runs through this in the batch manner a batch a group of of transactions and what it does is it just makes sure that the balances are updated for every single customers so you run them in the batch okay so it's a batch approach and that's what it's asking you for serves a batch approach so I guess in trying to answer that batch approaches you could check this out because I'm actually looking asking looking for what are the differences between batch and real real time essentially that's that's what I'm trying to get at at this stage there are differences in in the terms of a batch testing approach where all the data is sort of get it together and run at the end of the day which is a simple example there are issues there such as such as the time it takes to to I guess get it up all these these data to get up it's called it's called latency which is time related because of that it can be slow so the speed is slowed down perhaps the data are collected over time so it may not be run at the end of the day it could be

run at the end of the week for example all right so those are the sort of things I'm looking for
in terms of batch versus real time so real time is on the fly so obviously on the fly it's you know
for a customer for example that is more the preferred method because then you don't have that
latency you don't have the time you you know your balances are updated like straight away the speed
tends to tends to be faster those sort of things that are you know that are considered in more of a
real time basis so I guess when you look at 5.3 I want you to look at perhaps going to Google
mode or chat GPT where you're question or your problem is more like what are the differences between
batch and real time in terms of testing big data okay that is probably something that you would
kind of look at and and learn from what they would give you as as as a response the differences
between batch and real time testing okay so on then note that steps where we are this is the
number six so this is more like you sit down and work on this if this does not this looks like
Greeks to you or Italian to you or even Chinese to you I listened to last week's
workshop when we had it last week towards the end I actually went through this same exercise
more or less you went through some of the same question but just in case you you may not want to do that
what I have done is I have begin going back to my uh going back to the the learning side I have put
a under introduction to big data cluster if you click that what I've done is I've added in this
this PDFs it's a lab on on pandas all right obviously we we started talking about pandas we started
talking about data frames and a number of these questions here let me just click on it
there you go so you can go through this in your own time this is exactly what we did last week so
this is more on a a documented view of panda for example why we are working with pandas because pandas
working with structured data it is more of making reference to data frames which we talked about
which are basically rows and columns all right and then we started looking at importing this so this
is a a python command and we went through these things here all right so again these are the things
that you will need to go through in trying to address question number six so go through this
these are the commands we talked about sub-setting data we wanted to look at only eight out of that
if you can remember how many columns that we have in the original data set was like 70 plus is huge
we didn't want to have all that rubbish we wanted to make sure it was clean we selected eight of them
but how do you do that in python all right so go through these exercise because that's exactly
what questions six all about you can make reference to these of course these courts here
and then we did a couple of simple exercises around sub-setting data according to the time
so if we just wanted time to be bigger than in a larger or equal to the 1900s then that's the
way to do it okay so exploring data a bit about the descriptive stats in there all right so you can
use some of this to to help you to aid you to assist you in answering question number six and
that's that's the end of that okay so that's that's in this part here and obviously you can
download and keep it or whatever you want to do within okay so that's what I have let's go back
to this part here so that that's going to address those question number six okay what time do we have
oh nice almost on the door one minute to spare so anyway not have any questions related to this
this assessment because next week we're going to go on to assessment task number two all right so
as you can see this it gets you know it can get quite overwhelming if you tend to fall behind
okay any questions Eric Vlanda angry at Lisa Automas at this stage before I let you go
all right sounds good if you do have any questions between now and next week
so I'm having trouble let me just go back to what Ingrid has said I'm having trouble assessing recordings
is that is is that because you don't know how to do it in grid or is it you actually have
problems going to it no no I'm having troubles I'm looking at it I've been able to look at them
and play them in the past but today when I go to view recordings it just gets a little circle
spinning around and nothing happens okay let's let's try that so this is your this is your
learn line right yep and I have to simulate a student because I'm not a student so it tends to be
a bit strange so what you do is you you you click here I would assume
sure and view our recordings is that what you will you will get so I click on that so I'm I'm
I'm under a student view so I get I get this view here yes you're I was getting that but today I'm

just getting it stuck on that first little spinning circle right okay not gliding them for some reason maybe yeah trust me Ingrid it's it's there sometimes it's it's I don't know internet or because I yeah because it when it starts to spin around what it says is actually it's hanging and it doesn't maybe it's it's it's waiting for something I'm not exactly sure but it shouldn't be spinning forever you should get to this and once you get to this it's just a matter of clicking I've been accessing them the whole time it's just today so I'm not sure what's different interesting if it if it's if it's not on tomorrow it could it could be I don't it's kind of hard to pinpoint and trouble should because you know the you change your laptop for example or do you use a different system and to to get access to the to the recording but the recordings are there so you should have does anyone else have problems
looking at my recordings so Thomas said today CDU took a long time so maybe it's we are having some some IT issues for today finger crossed Eric said yeah you can you can you can download the recording if you want to yeah you can do that so it depends again on if it's a large one is going to take a while you need because can make the recordings it should be isn't it isn't it not available Eric I thought I I did set it to be you can download okay all right let me check on that Eric thank thank thanks for that I I usually do do allow my students to download this there's nothing to nothing to no no dramas there let me check on that and I'll fix that if I have to yeah that will be good all right I'll do that all right so on that note thank you very much guys and I'll see you next week same day same time all right cheers bye

0

Lecture 7 - XBD402 T3 20230906 1631 Tutorial 2 AT2

Okay, great. All right. I've got Lisa and Glenda. Welcome. So I was just talking to Lisa about marking the first assessments. I've marked a few already, but I'll just make sure that I'll mark everyone's submission for the first one. Not the first one, the first unit. 81 and 82 of the first unit. All right. By this week, so you have an indication. It should be pretty straightforward at this stage. Okay. So today, Glenda, do you want to say anything? No, it's fine. I'm nearly done with the first one and the second unit. So I'll grab something from there tomorrow. Okay, great. No, that'd be good. I'll try to mark it as soft as soon as I can because, yeah, it's quite a number of students. So got to get ahead. All right. So I've got Ingrid that have come on board. Welcome. Good afternoon. So today, we are going to look at the second unit, second assessment. The second unit has been probably, I mean, just to be honest, of all the four units, I think the second one would be probably the most, I wouldn't want to use the word boring, but it's very, we talked about a number of things, but it's only when you get to work and play around tools that it really makes sense to you. At this stage is just very much a conceptual discussion around testing. But what we did actually in the first week, which was the lecture itself, was we actually went through some hands-on cleaning some data and all that. So I think that's more of a better value than talking about some of these testing tools. We are as an industry not really up to that speed yet. A lot of organizations are not using testing tools. A lot of the testing strategies, especially around big data hasn't been formed yet. So a lot of this were, the whole topic of testing is really from testing systems. So organization brings in a new system, can be a payroll system, could be some kind of a management system, and before it is, you know, it gets rolled out to the industry, it has to be tested. So they tend to follow the same kind of strategies in that sense. Okay, just a couple of home keeping stuff I just want to go through. I wrote this down, but I'm not sure. Okay, all right. So I think number one is if you want to go ahead and have a look at recording. Let's see, am I sharing the right? Yep. If I want to go and have a look at recording for now, let's say I click on this, this is a student view. Yep. View of my recordings. And you'll probably see something like this. All right, where the latest one was, yes, we did talk about 402.81. That was our last recording. This is all correct. But going down over here, we find that the first lecture has disappeared, all right, which is 401 lecture, which is here somewhere, has disappeared for some reason. So it does happen because sometimes learn line, or rather this particular tool within learn line tends to filter it off. So to be able to get it back here, this is what you would need to do, recent recordings. Make sure that you actually go back to, say what it does is it has today's date, but it's from today's date as well. So it's best if you go back to, oh, when did we start this? Maybe the first of August. Yep, the first of August will be a good start. And, and, yep, and behold, what happens then is you have the lectures coming up here. Okay, so that's just a little trick to know because towards the end, students start to panic and say what happened to all the earlier recordings because I need to listen to them. All right, so that's that bit over there. So let me close that down. The other thing I wanted to just say is if you were to go to, I have managed to put in a recording here. So introduction, big data cluster, if you click on that, if you go all the way down, I've added sort of a slightly more than 10 minutes, 12, 12 and a half minutes of content, which is really a brief conceptual view of Hadoop. Now, Hadoop is a really complicated open source platform. And again, you can use Hadoop, people use Hadoop as a framework to in big data. So those of you who are actually directly involved in the management of huge amount of data would definitely be looking at Hadoop as a platform. So basically what I've done is, and I've spoken on Hadoop a couple of times, but I really never explain it. So what I did was I sort of put together some slides and just gone through some of these things. So hopefully it's easy to understand. I try to make it sort of a Hadoop 101 kind of perspective, so that people can understand it even from the street,

hopefully. Yeah, so that's what I did. And it's up here, you can click on it and there's a video. What I did not mention in that clip is the fact that, and I kind of put it down here, is the fact that there are four components of the Hadoop ecosystem. So this will make more sense once you have listened to the video. And then you'll say, well, what's so special about Hadoop? Well, it works within sort of an ecosystem, a very complex ecosystem. It has its own file management system. It has its own distributed storage mechanism. It has its own resource management system and so on. So I've just put some over here. It has its own programming model. And you may have seen this word many times in the modules as you read through them, map reduce. So map reduce is just a, you can look at it as a programming tool within this ecosystem, this Hadoop ecosystem. All right. And it's used for data processing within that ecosystem. So have a listen to this 12 minutes clip. And then after that, hopefully then it'll make a little bit more sense. But today we want to really want to focus a bit on what else? Let me just blow this down. 402 is what we're looking at today. And the resources are there. Okay. So this is the second 82 of 402. And I've set the time properly. So now it's due on that and just the midnight. Okay. Let's see. I think that's all I want to say in terms of the the housekeeping stuff. Does any one of you have any questions before I actually start this second assessment? We've got Thomas on board now. So Glenda, Ingrid, Lisa and Thomas, welcome. Right. So let's go through them.

I've actually went through and reduced the number of questions from the original set of questions that was posted weeks ago. So this is the new version of it. All right. So keeping in mind, we're looking at big data testing. And activity number one sort of makes you go back to the first unit here. So XBBD 401 was the first unit. This should say 82. So activities three and four. And basically, I'm not sure if you still remember that particular assessment, 82, there were two somewhat related activities in there. And they had to do with, let me just pull it up because I think I got it up here. There you go. 401. I was just looking at it. So this is activity number three. And further down is activity number six. Okay. Both of them are related. So the third one, just to refresh your mind, the third one is just, it's fairly simple. It's just asking you to reflect in the context of the case study, which is the e-commerce organization. And that's our case study. It wants you to sort of put down here in these boxes, the type of data that you're going to capture. So in a typical transactional data, what are the sort of data that you're going to be capturing for this particular organization, for this particular case? Is it going to be sort of a customer transactional sort of data? You will also be looking at a non-transactional one and the type of data that you'll be capturing. So it's just a sort of a text-based answer over here, just describing the sort of data set that you will be collecting. And then if you look at activity number six, which is related to it, activity number six is more into you generating the data sets. And I think I went through the process where you would go into chat GPT and you will, on the PROM level, you would just type in some instructions for chat GPT to generate a typical sort of a transactional data according to the fields that you want.

Okay, so that's that particular assessment. Now, assuming that's done, oops, wrong one, sorry. So assuming that you know what's that all about now, this is a fairly straightforward. All you need to do is explain the two data sets in the context of why you have chosen them with their respective fields and variables. So it's basically a sort of a repeat of what was asked, but that's just to set the scene for this particular series of questions here. All right, so the keyword is explained. So you can explain it fairly briefly at this stage, saying the first transactional data set that you have thought about as a data analyst should look something like this. All right, A and B, A would be transactional, B would be non-transactional. And that's what I'm looking for, something really basic, just to refresh, to review exactly what you have looked at. In this particular assessment, just make this a little bit bigger. Okay, so that's number one. Number two is about

legislation. So legislation and then going have a look at the organizational policies and procedures. And the question here is what are some of the organizational policies and procedures when it comes to obtaining raw big data for testing? So for this, what you need to go back to is really the case study. So I'm just going to pull up the case study like so over here and run through it fairly quickly and find out exactly whether I can actually get that sort of information from the case study as in some of the things that's related to legislation or some kind of a strategy. And if you go down here, well, there's some things around privacy and all that, but it's still not quite related to what we want. What we want is, going back to the question here is when you are trying to obtain some data for testing. And I kind of highlighted it here to make your job a little bit easier because this question can be approached at a very high level, very broad perspective. And behold, what you see here is some kind of an organizational procedure. So this is the kind of policies and procedure that was, I guess, for the last couple of years, that was, I guess, formulated by this particular organization. And they say that we have developed a seven-step process to help us to look at this strategy. And what you can do is to just look at this seven steps and just extract those seven steps from here and just put it into the space over there. So this will be step number one here, identify what you want. Number two, leverage on a proven data strategy. I know that it doesn't really answer the question directly. The question asks for a big, I'm trying to read it, big data for testing. All right. So, but these are really broad sort of strategies by the organization. So that's good enough. It doesn't really give you exact information around testing. So it's fair enough. So if you don't have that information, that's fine. Just copy the seven highlighted areas as the seven sort of points, seven steps. And all you need to do for this particular activity is just to put them down here as one, two, three, four, five, six, seven. Okay. It's not too hard there. Okay. So it's more of obtaining some big data for the organization as opposed to very specific here for testing, but that's okay. So I did go through the case study. There wasn't anything that's very specific for testing. Okay. So that's activity number two.

Let's move on unless you have any questions.

The third activity is around validation. Okay. So this is where you validate your data from various sources. Okay. So to do so, what we need to do is it sort of reminds us what the kind of steps are to do some validation. If you don't like the word validation, it's about basically making sure that you have the right, it's all nice and clean. That's what it's all about. All right. To make it all nice and clean, you need to do some validation. So it talks about some of the steps here. And this is pretty straightforward. Yeah. You need to know what your sample is to test, because you can't test the whole data set. If indeed this was a large big data, it could go into the millions. So we usually look at a sample, which is fine. And then we can do the testing on that. Validate. Yep. We can validate and we can validate some of the formats in there. Okay. Things that we kind of spoke about. So in terms of completing this particular part of the assessment, it's all about filling in these little boxes on the right-hand side. And again, I'm looking at just one or two sentences. I'm not looking for a huge massive essay on, for example, how do I determine data samples? You can put something down like, for example, over here, if you have a large volume of data, you'll probably want to validate a sample of the data rather than the entire set. Something like that. It's nice and simple. It sort of explains probably in a sentence exactly what this is all about. And that's what I need at this stage. Nothing more than that. That will suffice. In terms of validation of the data of the database, you can put things like, we just have to make sure that to validate the database, we want to make sure that the database has all the required information and data in your existing database. All right. It's a pretty straightforward process. Now, if you want to describe the process a little bit more, like validation, why validation of databases is important, you can indeed go ahead and write a couple of one or two sentences about that. But basically, this is just a basic understanding. Some of you might look at this as definitional. So what does it mean when someone says, I need

to determine a data sample? Where do you get that from? And so on. So there's a couple of ways to answer these sort of questions. That's what I'm trying to say. And the third one is a little bit more specific. It talks about validation of data format. Again, when you're trying to pull things down, like what we're saying here, we're pulling information from various sources to ensure data quality. We want to make sure that our data formats are the same, such as my favorite example, male gender. So gender is male or female. Some data sets will put it at zero and ones. Some could be the reverse, one and zero. You need to make sure that those are all consistent. So make sure that there's no incorrect different kind of formats when you start to integrate all this together. Okay. Scripting. Now, this is where we use other methods, I guess, to validate information. The scripting here refers to actually the scripting that can be performed by a, I'll just write this down, by a scripting language, such as Python. Okay. So you can use Python, for example, as a scripting language to do validation. And you saw some of those examples when we started to, you can test for missing values, you can test for duplications and so on. By the way, the next unit would be more, I think it's a little bit more interesting because the next unit, 403, is about analysis. So it's very heavy on using Python in doing the analysis, unlike testing, which is pretty much a conceptual perspective at this point. But the next two units, as in analysis, and then the last unit, which is on plotting and visualization and all that kind of cool stuff and on bar charts and all that, those are definitely very much Python-centric and obviously Excel-centric as well. So we'll be using that tool a fair bit.

So scripting language such as Python is good enough. So, and you can say, well, it's used to, you can write scripts for validation purposes. All right. That's about all I need to know at this stage. We haven't played around with enterprise tools to actually do big data testing or validation, sorry, validation. Those tools can be very complex and they're very proprietary because there are many different tools around. But basically, you can write here, you can use, you know, there are numerous enterprise tools around, which can help to perform data validation. I'm not sure if I put some examples. Oh, I did. Okay. So I did actually put some examples here of some of those tools, enterprise tools that can help in validation. In fact, if I just cut that off, oops, cut that off and just focus a little bit on MatReduce. So MatReduce is a really one of the most common tool that we have in the industry. Some call it a programming tool, but it really helps in the world of Hadoop. And that's why I sort of talked about Hadoop in terms of data processing. Okay. So you need to understand Hadoop before you can understand what I'm saying where a tool like MatReduce can help in processing of data. Or also keep that in mind. You can cut and paste MatReduce and put it down here. An example of an enterprise tool in the Hadoop ecosystem is MatReduce. Okay. All right. Okay. Let's have some, at least some participation at this stage here. I kind of modified this question and put it into a more, I guess, high level perspective. And it is about metrics. So big data metrics, testing metrics. All right. And the fact that I said, well, there are two categories of such metrics. Metrics are just measures. Some of you might want to look at metrics as KPIs, key performance indicators. So there are process metrics to measure how effective the process is. And of course there are KPIs or metrics around how effective or efficient the project is. All right. So these are big data projects. So as I said here, some of the process metrics around efficiency, effectiveness, and productivity could be of the testing team. In terms of the project metrics, it could be the cost of the project. Did it finish on time, the schedule, scope, risks, and so on and so forth. So over here, the question itself is fairly straightforward. I'm asking students to give me some examples of these metrics. All right. I've given you, I've given some clues over here, but I just need to have actual names of some of these metrics. For example, one of the metrics that I use in determining, I guess, this particular training program is the number of students, the percentage of students that have attained a pass or have attained a CA, which is a competency attain. All right. That's my metric. So in the world of big data, they have their own metrics. So can anyone think of

an example of a process metric at this stage?
Error rate. Yep. Acceptable. I can handle that one. All right. So thanks for that.
Let me just go back to see. I think that was, where's my little button here.
That was angry. I was just checking on who it was, making sure. Angry. Thanks for that. All right.
Anyone else? It could be something that's more around, well, I mean, these things are not easy.
So let me just give you some examples. And I want you to sort of, if you can think about them,
or you can Google them to get more information, that's fine. We're all learning in this whole
process. One that's quite common is the test cycle time. Let me just put these things down.
I've got to type the test cycle time. It's really sometimes the testing itself comes from,
let me just put a dot here. You know, we have this planning process
from planning to execution of the test itself. We do the test and then we report on the test results.
This is a typical sort of a sequence of a process here from planning to reporting.
And what we're doing is we're just measuring the time cycle from start to end. How long does that
take? Okay. So that's a typical, very straightforward example of a process metric, besides the one given
by Ingrid, which is more on the error. Okay. There's another one which is called the,
I guess, what we call the execution rate. In other words, the number of test cases per unit of time.
All right. So I'm not going to put that down. Rate of output to input. Yes, very generic, but
and you can be generic at this stage. And I totally understand a lot of us do not have
backgrounds in big data or even any kind of analytics. So I'm pretty open to more generic
discussion around relating to what you have been doing in different industries and so on.
Now that's fine. Now those are good. Thank you for the rate of output to input.
Let's look at the project metrics. I think this hopefully becomes easier because we can put things
like as they've given us the cost. So we have the test budget. So this becomes much easier to
understand as opposed to a process metrics or project metrics. We can measure around test
budget, which is the costing anyone else that we can look at some of these metrics related to a
big data project of testing. So we look at costing, we can look at ROI. Yes. Yes. Yep. That gets a
bit technical, but that's in project management. That's fine. You can also look at the clues are
here. So it talks about calling. So that's a budget schedule is what schedule is.
If you want to look at it another way, it's the timeline. That's a measure. Yep. Great.
Yeah. That's enough. You don't have to list me a whole bunch of metrics, but you know,
as long as you get the idea, that's fairly good enough. Yep. Yep. Good. So those are the metrics
activity five. Let's get on to that. It looks at a different kind of topic. So as you can see,
the kind of different sub topics here, but they're all related to some kind of a,
the sort of the testing sort of the phase that we're in now. So over here, we're looking at the
whole idea of aggregation and segregation, which are opposites. So aggregate is to get together,
to combine and segregate is to divide and sort of more of a diverse approach to things,
split you up, you know, segregate. So what it's talking about here is the aggregation of data sets
over here. And segregation is talking about, you know,
putting it up into different data sets. Okay. Well, they kind of use the word here,
data into clusters, it means groups of clusters. Okay. So with the kind of basic understanding
as to what the aggregate is, when you put it together. So that's where you get the data,
which is yes, you need to gather them and then you need to combine them. I understand that.
And then you want to make sure that your combined data is accurate. Okay. That makes sense. And then
you look at segregation here, which is to divide is to pick up on all these different
data sets that you want to split into. Maybe we want to split into you've got one data set
and you want to split that into the males and the females. You want to split that into those who are
earning income more than 50,000 versus those that are earning less than 50,000. So you're
already thinking of the how you're going to segregate, you know, a single data set at this
stage. You can separate them based on transactional and non-transactional. You can call them as clusters,
you know, the male clusters and the female clusters, for example. So over here, what

I'm asking you is again, pretty straightforward is why perform? Why? I guess that's the keyword. Why would you as a data analyst perform aggregation? And I think that's a pretty obvious answer. And I think that's something that I spoke about because it's all about what? It's all about here, which is about combining data from different sources to provide maybe a summary of that data. Okay. That's all you need to put down at that little space here. Provide some examples. Well, again, many examples of why you would want to do that. Anyone, if you were thinking about aggregation now, why would you? And looking for a really good example here where you could, what kind of data, you know, will you integrate, put it together so it makes sense? What kind of data? Examples of those. Obviously data from different
different sets. But what's the common, what's the common denominator? What's the one? Let me give you one example so that, you know, your mind starts to combine demographics with geographical behaviour data. Oh, okay. Yeah, okay. I get what you're saying. I would interpret that as demographics. Yes, you can. But the word geographical tends to resonate more with me as in terms of an example. So I would be sort of maybe grouping postal codes, right? So I got different postal codes. Let's survey from different postal codes. And what I'm trying to do is to integrate these postal codes into maybe into a district or a region. Makes sense? In terms of demographics is kind of same story, right? So you might be wanting to integrate or aggregate people from different social economic backgrounds, for example. Yeah, those are the sort of things I'm looking for. On a business side is more of looking at customer data as an example, customer data from different departments. So, you know, I've got different departments selling different products, for example, you know, department A, B and C. What I'm trying to do as a sales manager is try to aggregate all these information and put into one single data set. So that's another example.
On the other side of the coin over here, we're looking at segregation,
right? So segregation, as we talked about, why do this? Again, the definition part of it is a process of segregating or separating data into different groups, categories or clusters.
Okay, that's the definitional. But again, looking at some examples,
which one would make sense? Why would you want to, you know, separate?
And again, I'm looking for some examples as to what you're going to set, what kind of data would you separate? You could be looking at the data that you're going to separate.
I guess the opposite side, isn't it? Income. Yes, yep. You could be trying to, you know, separate the customer data set based on financials such as income, yeah, age, minor, adults, yes, yep.
Yep, into gender, for example, into age, yeah, exactly. So I think you get the idea what these things are all about. Pretty straightforward. I mean, these are really basic but fundamental concepts in big data. Okay, well then, let's move along.
TV number six, I think this is a really weird one. I was trying to make it nice and easy for you, for the students, simply because you didn't have the hands-on for performance testing. Okay, what I've done here is, again, kind of put in boxes, you know, it's up to you to kind of move this thing if you don't like, you know, you want to, oops, what did I do? Change everything. Don't do that. Okay, in terms of writing a little bit of anopsis in each of these little
in terms of writing a little bit of anopsis in each of these little boxes here. But what I've given you here is this part here. I've given you the answer. So the answer is there. So don't do anything about it. The answer is given, or, you know, one option is. So it's given down here in this form. And it was the whole idea was this space here was blank.
But because I think in real life, if you have not done a performance test before, this would make it really hard to work on. But so what I've done here is presented a typical performance testing for data. Okay, and again, performing testing in terms of this particular case, you have to go back to I think back to this case study. And it does talk about
not in this case.
Doesn't talk about sorry, it doesn't talk about performance testing, talk about performance management, which is slightly kind of different. Okay, ignore what I've just said. So let me just

go back to here. So this is a typical performance testing process. And it's not a very, very general process. So what you need to do is let's set up the application. So, you know, you can, depending on the system you're using, again, these are all there are many different testing systems out there. Let's say you pick one of those, you start here, you start the process. And you can see that there are a lot of different testing systems out there. So let's say you pick one of those, you start here, you start the process, you identify what you want to test on in terms of the performance, it uses the uses the word workload, which is something that I kind of use also in my in my video on on Hadoop. It's a Hadoop sort of terminology which says you can tend to split up all your work into workloads. Okay, so you can have a large workload, and you split into workload number one, two, three, four. So it's step two is about trying to identify all the different workloads and the sub workloads that you are trying to test in terms of the performance. Who are these for? Individual clients that are doing the testing. And then you do the testing itself, the execution and analysis and then it comes up with the optimal, you come up, basically this is the report itself. Okay, sometimes the objectives are not met because you have a target. So you have a target performance targets. And when you do the tests on the data is not meeting those targets. Therefore, you have this thing here that says objectives are not met. Okay, so this is just an approach, one of the approaches for performance testing.

As I said, the answer is given. So don't do anything about it at this stage, but just, you know, but you may want to know a little bit more about it. Obviously, you can go into all the other tools that we've been playing around with Google and chat GPT and all that stuff to read up more on this. But this is, I would I would feel that this is beyond the kind of scope of what we are trying to achieve in this particular unit here. But in terms of the why we conduct performance testing, I think that's something that we would, yes, is something that we would need to know as data analysts in a very broad perspective. So in terms of ingestion and throughput, getting ready for staging, for example, is ingestion. So you're pulling in all the data from all the different sources and you're putting into a data warehouse, for example. There are performance testing that you need to work on, such as if you think about this, how fast can the system actually pull all this individual data from all the different various data sources? It's an issue, right, related to performance testing. Let me say that again. So in terms of performance testing, where the scenario is you have, let's say, five different data sources. One, two, three, four, five. One could be, you know, one could be a text base, one could be an Excel, one could be an Oracle database. And what you're trying to do is pull in all the relevant information into a into a data warehouse. All right. So what are some of the criteria? What are some of the parameters that you'd be thinking about when you're looking at such a thing called performance testing? Hopefully, the thing that comes to your mind is, is it possible? How fast can this thing actually pull those information out and put it into the data warehouse? So these things that you'll be thinking about, you know, the volume of the data, the speed of the data, and so on. And some of these answers here are also related to the one at the bottom here, which is data processing. So in terms of processing the data itself, there's also some questions again, how fast it is, and also looking at the speed, which is about how fast it is, how often. Those are so, so in your research, I would be, there'll be one of the good things to look at to search is searching for some of the criteria in conducting data performance testing.

Yep. Try that. Try that out. Put into chat GPT. What are some of the, of course, I think you need to post in some form of a question. So what are some of the search criteria in conducting, or what are some of the considerations in conducting big data performance testing? All right. And that will give you an indication of what to put into these two boxes here. All right. How was that? Anyone have any questions related to this one? It's a strange one, as I said, but you know, it's doable.

Okay. I'm looking at my time. So I think we are on track, but let me just go a little bit quicker.
All right. Activity number seven. So this is where we have talked about last week,
where we talked about batch, batch processing in terms of testing and real time. We know what real time is, right? Real time is like going to an ATM and looking at your balance in your bank account and withdrawing the money and straight away the balance gets updated real time. That's real time. That just sort of keeps all the transactions together and at the end of the day, perhaps you'll then run it through one shot. All right. So you will not have a very accurate balance until maybe the next morning dispatch. So how do I actually write these things down? Again, you know, over here, it's kind of silly, but I would say things like involve test procedures that would run the data in batch processing mode, for example. Perhaps then explain a little bit of what batch data processing really is all about. That's a good product that does that. Spark.
You can read more on Spark as a product to do batch processing. Real time, again, I think is pretty obvious. Real time is real time environment. Talk about the real time environment and it's, and people use that because it's stable. If you have a stable banking system, for example, and they're very confident that we can do this real time, we can have the latest balance and we're very confident that that value is correct. It's all about confidence. It's all about stability.
So that's that bit over there. Fairly simple. Activity number eight requires you to isolate substandard data. So this is where you start to look at substandard data. I mean, data that are not normal. You know, an example of some of these are over here. You know, data that is incorrectly transcribed. Maybe an extra zero. Maybe the decimal point has moved one or two, you know, to the left and right, that sort of things. What you need to do here is to go into each of the boxes and sort of talk about, for example, let's look at this one. Data on scrap paper, right?
So this is where you have data that's manually written
on files and you're storing, I guess, maybe photos of these data that's written on just paper and someone has just somehow uploaded that information onto the system. All right. So it's about what you think about that, all right? How do you do that? How do you do that? How do you do that?
All right. So it's about what you think about that, all right? Some of, you know,
how do you actually manage that? Let's go to something which I think we have done before and so that you'll make more sense. Incorrectly transcribed data, all right? So in terms,
instead of typing in a temperature in centigrade, I type in a temperature in Fahrenheit.
Okay, that's incorrect, conscribed. So the data that's wrongly typed
because of typos and things like that. So that's why I'm looking for something generic,
some statement about why, you know, why sometimes this happens and some examples of those things such as, as I said, temperature, maybe typo instead of 0.05, I put 0.5, that kind of things.
And some examples of those, a few of those would be fine. Inappropriate error repairs.
This is a strange one. I was trying to scratch my head what this means, but it was about really data that contains, let me just put this down,
that contains useless repairs. And they use the word repairs in terms of correcting. So their way of looking at useless correction. Right, so in other words, are there some data,
and we sort of played around with this when we were looking at Python, remember, we were looking at data and so we were trying to correct some of those data. But there are some data in certain data sets that are so useless that you don't even want to repair them. And I think this is what it's trying to say, data that you don't even want to repair them because maybe they are irrelevant, you know, and it's something that is not been, it's not something that you will need. So you want to just remove that column, for example. So those are the sort of examples of useless error repairs. You know their errors, but you don't even care about repairing them.
Okay, that's that part over there. And the last one is data that's inconsistent with protocols.
It's really about, with protocols, again, somewhat misleading, but what it's trying to say is inconsistent with organizational policies and procedures. I think that's what it means.
Two more to go. Okay.

Okay. Okay, this one is about general storing results of validation activities
and giving evidence according to policies. Okay. Again, case study, create a test report,
store the data. Okay, so this is about creating a test report. So the results of a test.
So you run a big data test and you have a kind of report, maybe it's a performance testing report.
The other thing is also under consideration is where do you store this. So it's a pretty generic
question here. Oh, look here. Identify some legislative requirements
in terms of storing test results. I'm not sure what that means, but I guess this is looking at maybe
there's some legislation related to real time when you're monitoring data on a real time basis.
I'm just speaking aloud here. Perhaps it's looking at to store.
Looking at to store.
I'm just highlighting things that we need to focus on to securely store that.
Because when I'm thinking about storing stuff and making sure that there are, according to
legislation requirements, and then because I'm teaching cyber security, I tend to approach this
question in sort of a cyber security mindset. And I'll be putting in there such as making sure that
there's anti-virus protection to make sure that it's updated, making sure that there is,
you know, the systems that I'm storing this test reports in have intrusion detection and
preventive systems. So it's a very cyber security management perspective here.
The other perspective, that kind of mindset I have is also the encryption of data itself.
So that people can, if they do break in and they do steal the data, it's all encrypted so they can't
read it. All right, so that's what I would put in. So I guess an answer to this would be some of
those things that I've just said in bullet points such as anti-virus protection,
intrusion detection, and prevention systems, and also encrypted data. Okay. Very simple.
Very much a, put a sort of a cyber security perspective to this
management perspective. Now you may have perspective. You may have a different slant to this,
which is fine. Happy to read what you have in terms of how you would address this particular one. But
at the moment, my mindset is just focusing on that.
Number 10, which I believe is the last one. All right. Yep, it is the last one.
It is about cleansing. Okay. So something that we played around. Okay. And we did this stuff like
when you clean your dirty data set, you tend to remove irrelevant data. You tend to make,
get rid of missing values. It's over here. We played around with that. You get rid of duplication.
Typo errors like 0.5 is 0.005. That sort of thing. Convert data types.
Okay. All right. So what it wants us to do is to, for each one of this, I guess,
explain. All right. So let's go through this so that we all have a common understanding
to remove irrelevant values. All right. So to me, that means I want to remove useless pieces of data.
That's how I would approach it. Remove useless pieces of data because it's irrelevant.
And that's good enough for me. If you put that down here, get rid of duplicate values.
Why? Because duplication is, explain why duplication is useless. All right. Why do you,
why would you want duplications in your data set? That's why you need to get rid of them.
Okay. To avoid typos and similar errors. Well, what can you put here? Well, you can put, say that
it is, sometimes it's unavoidable because it's classified as a human error. Okay. And it's really
hard to spot, isn't it? 0.5 is pretty close to 0.05. And visually, if you're looking at hundreds
and hundreds of rows of data, you may not even spot the difference between 0.5 and 0.05,
to be honest. So you need some automation in that process and therefore you'll be using
sophisticated testing tools to do that. Convert data type is in reference to making sure that
everything is uniform. That's how I would look at it. Data types have to be uniform. So in other
words, if my data type for this particular column is numeric, so these are numbers,
they have to be numbers. So you cannot enter, for example, strings of data.
Okay. Because sometimes you can say, you can type in one, two, three,
the systems look at it as a number. A number is one, two, three. You can even convert that one,
two, three to a string of characters. So the system can look at it as a character,

a string, which is consisting of one, two, three, not numerics anymore. So we want to make sure that in terms of cleansing, we want to make sure that the input, the information that you see within those cells are indeed of the same type, the same data type. And of course, the last one is taking care of missing values. We went through some examples of those in terms of how to handle missing values. All right. And things like, you can replace them with NA. You can, in the case of Python, I think it replaces it with, I don't remember, NIA or NNA or something like that.
All right. So different systems look at it quite differently. Or you can just leave it blank or put a value there, maybe zeros to represent all missing values. But again, it depends on the case. It depends on the circumstances. Oh, that was a fair bit. I thought it was...
Okay. So that's the... It's pretty straightforward, but there's a number of questions there.
At this stage, does anyone have any questions?
Does any of this, Manesta said I'm actually doing something, like doing an example. Was it just it's theoretical, isn't it? It's not... Yeah, it's very theoretical. Give me some examples of the theoretical, explain why. Like we don't actually have to do an exercise in Python or Excel or something. No, no, no. We can't do this because we just don't have the tools.
Yeah, sure.
Yeah.
Yeah. That's why I was saying that this is the only unit that doesn't have a lot of...
Yeah. So you don't have to give me examples that you have done in Python, for example. The kind of exercises that we went through like two weeks ago, you don't have to do any of those. It's just... Yeah, just... Yeah.
Theoretical kind of a brain dump, I guess, into these spaces here. Not very interesting, but hey, that's what this thing is. I might rethink about this next time when this is offered.
But for now, that's what we have. Anyone else have any questions?
No? I just look here. Yeah, nothing in the chat. Okay. So it's... Well, I have a minute over the hour, which is great, perfect. So if there isn't any other questions, I'll call it call it a day and wish you a good evening and we'll see you next week. So next week is going to be looking at a new unit. So at this point, we are halfway through in terms of, I guess, the delivery part of it. But next week, we'll definitely get into the analysis part, which I think is much more interesting than this stuff here. It's where we'll start to look at actual... We'll have to use Python to play around with these things. All right. So in that sense...
All right. Catch you next week. Same day, same time. All right. Thanks. Bye.