TOPIC A

INTRODUCTION TO DATA AND DATA ENGINEERING

INTRODUCTION: DATA AND DATA ENGINEERING

CONTENT

- Describe Data
- Describe the role of Data Engineer
- Explain how the data are generated
- Compare the data for transaction processing versus analytics processing
- Differentiate Structure data, Unstructured data, Semi-structured data

Reference: Barton Paulson, Big Data Foundations Concepts and Techniques, LinkedIn Learning

Data forms include text, stream, audio, video, and metadata. Data can be structured, unstructured, or aggregated. For structured databases, data architects define the structure (schema) as they create the data storage in platform technologies such as Azure SQL Database and Azure SQL Data Warehouse. For unstructured (NoSQL) databases, each data element can have its own schema at query time. Data can be stored as a file in Azure Blob storage or as NoSQL data in Azure Cosmos DB or Azure HDInsight.

Data engineers must maintain data systems that are accurate, highly secure, and constantly available. The systems must comply with applicable regulations such as GDPR (General Data Protection Regulation) and industry standards such as PCI DSS (Payment Card Industry Data Security Standard). International companies might also have special data requirements that conform to regional norms such as the local language and date format. Data in these systems can be located anywhere. It can be on-premises or in the cloud, and it can be processed either in real time or in a batch.

Source: https://docs.microsoft.com/en-us/learn/modules/evolving-world-of-data/2-data-abundance, 15 Nov 2019

Data engineer

Data engineers provision and set up data platform technologies that are on-premises and in the cloud. They manage and secure the flow of structured and unstructured data from multiple sources. The data platforms they use can include relational databases, nonrelational databases, data streams, and file stores. Data engineers also ensure that data services securely and seamlessly integrate with other data platform technologies or application services such as Azure Cognitive Services, Azure Search, or even bots.

The Azure data engineer focuses on data-related tasks in Azure. Primary responsibilities include using services and tools to ingest, egress, and transform data from multiple sources. Azure data engineers collaborate with business stakeholders to identify and meet data requirements. They design and implement solutions. They also manage, monitor, and ensure the security and privacy of data to satisfy business needs.

The role of data engineer is different from the role of a database administrator. A data engineer's scope of work goes well beyond looking after a database and the server where it's hosted. Data engineers must also get, ingest, transform, validate, and clean up data to meet business requirements. This process is called *data wrangling*.

A data engineer adds tremendous value to both business intelligence and data science projects. Data wrangling can consume a lot of time. When the data engineer wrangles data, projects move more quickly because data scientists can focus on their own areas of work.

Both database administrators and business intelligence professionals can easily transition to a data engineer role. They just need to learn the tools and technology that are used to process large amounts of data.

WHAT IS DATA? (DIKW MODEL)

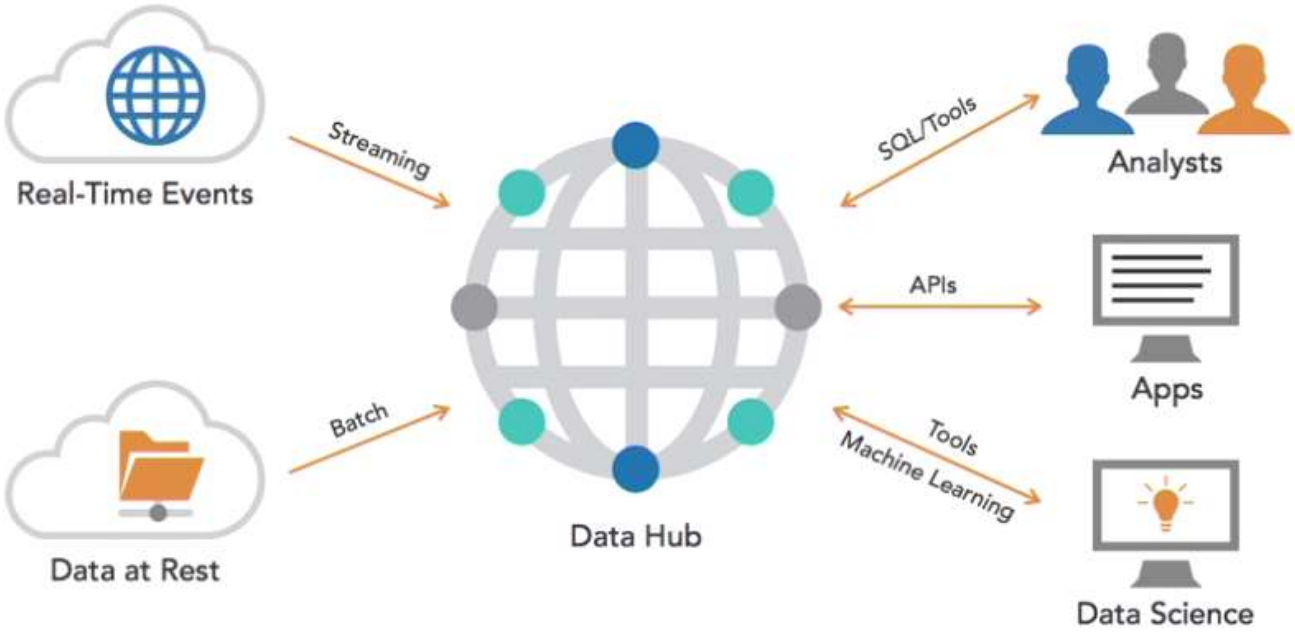
Use **mask**
Avoid Outdoors
Carry an umbrella

Singapore West weather in the afternoon and haze reading.

Singapore Island, West, Time:13:01:05PM, PSI 210 PM 2.5 ug/m3, Temperature 32, Humidity 62%

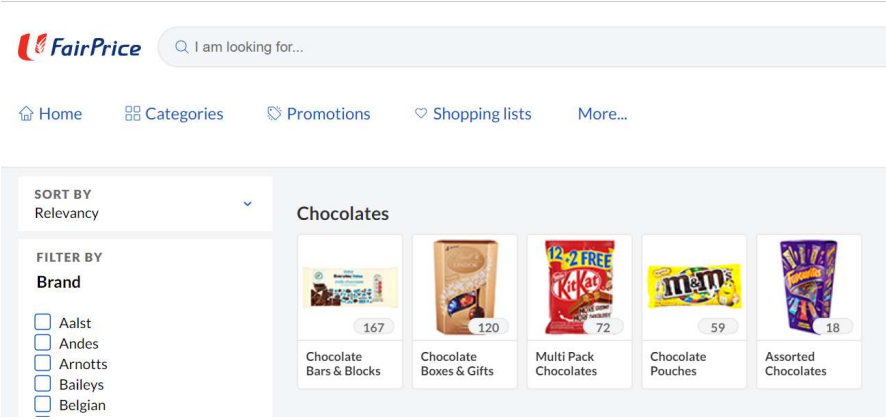
SG-W:13:01:05:210:32:62:C

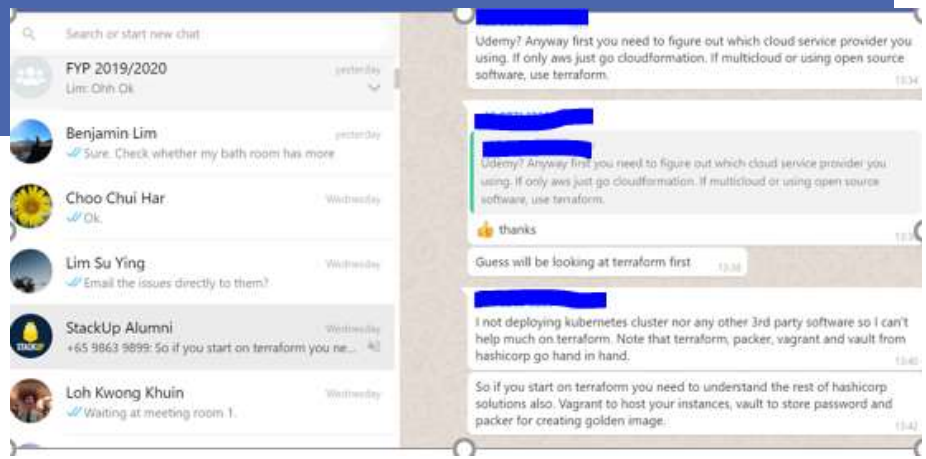
DATA ENGINEER ROLE

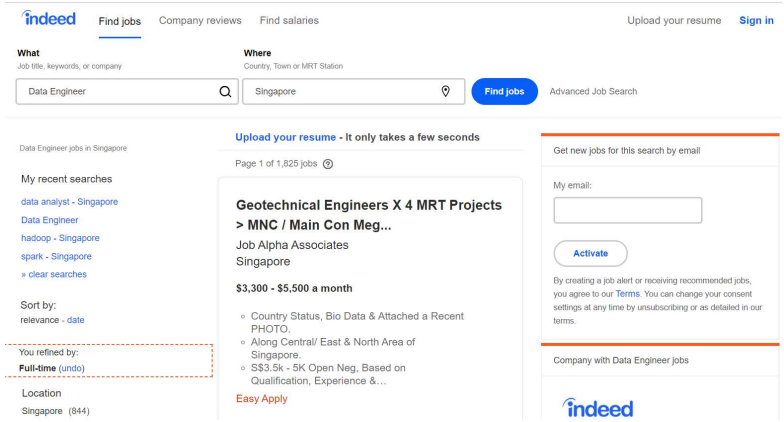


A

WHERE ARE THE DATA?







HUMAN-GENERATED DATA

- Payment (Transaction)
- Online Purchases (Transaction)
- Social Media Postings
- Text Messages
- Cell Phone calls
- ….

MACHINE-GENERATED DATA

- Cell Phone connect to towers
- Web Crawlers
- RFID Readings
- IoT data
- ….

OLTP & OLAP

Online Transaction Processing

- Handles recent operational data
- Size is smaller
- Goal is to perform day-to-day operations
- Use simple queries
- Faster processing speeds
- Required read/write operations

Online Analytics Processing

- Handles all historical data
- Size is larger
- Goal is to made decision from large data sources
- Use complex quires
- Slower processing speeds
- Requires read operations

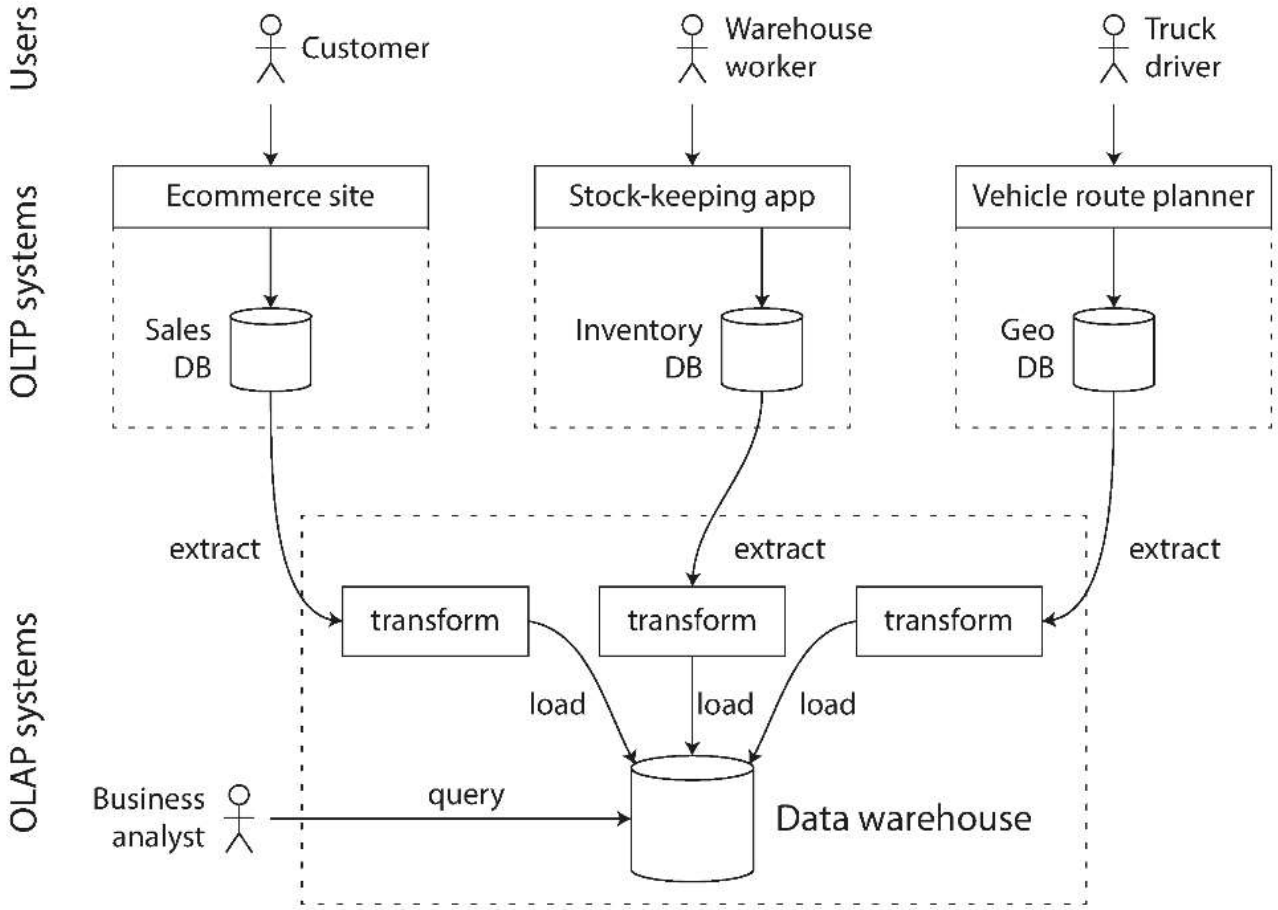
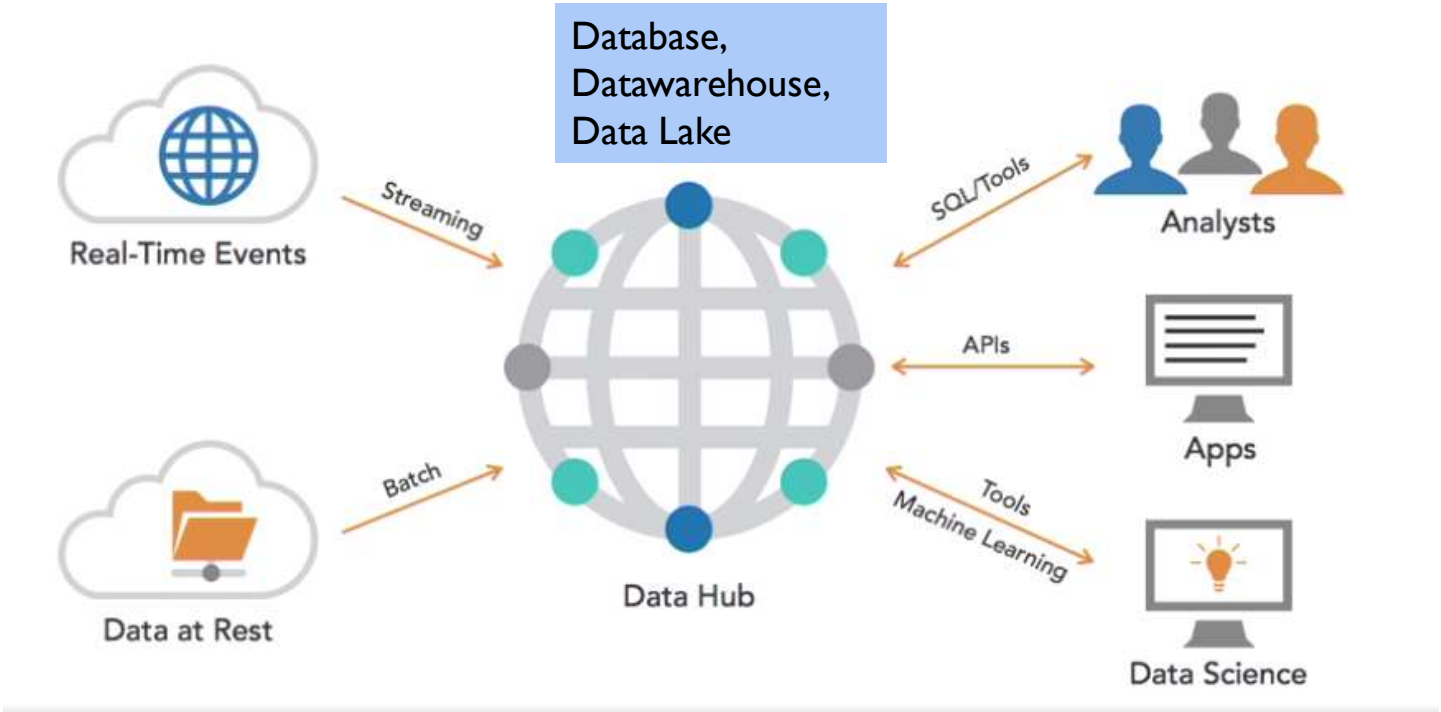


Figure 3-8. Simplified outline of ETL into a data warehouse.

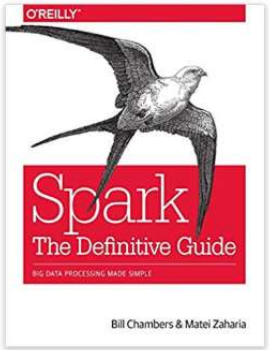
Source: Designing Data Intensive Application, Martin Kleppmann, O'relly

DATA ENGINEER ROLE



amazon.sg/Spark-Definitive-Guide-Bill-Chambers/dp/1491912219/ref=sr_1_fkmr0_1?keywords=Spark+SQL+definitive+guide&qid=157

You purchased this item on 2 Dec 2019.
View this order



See all 2 images

Spark - The Definitive Guide Paperback – 8 Mar 2018
by Bill Chambers (Author), Matei Zaharia (Author)
27 ratings

See all formats and editions

Paperback
S$38.26

5 New from S$38.26

Get it Sat, 28 Dec. with FREE delivery.
Get it Tomorrow if you order within **7 hrs and 24 mins** and choose faster shipping at checkout. Details

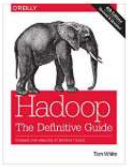
Learn how to use, deploy, and maintain Apache Spark with this comprehensive guide, written by the creators of the open-source cluster-computing framework. With an emphasis on improvements and new features in Spark 2.0, authors Bill Chambers and Matei Zaharia break down Spark topics into distinct sections, each with unique goals. You'll explore the basic operations and common functions of Spark's structured APIs, as well as Structured Streaming, a new high-level API for building end-to-end streaming applications. Developers and system administrators will learn the fundamentals of monitoring, tuning, and debugging Spark, and explore machine learning techniques and scenarios for employing MLlib,

Read more

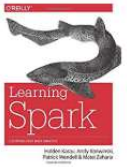
Report incorrect product information.

Customers who bought this item also bought

High Performance Spark
Holden Karau
15
Paperback
3 offers from S$46.48



Hadoop - The Definitive Guide 4e
Tom White
75
Paperback
S$48.42 prime



Learning Spark: Lightning-Fast Big Data Analysis
Holden Karau
74
Paperback
S$44.67 prime

Programming in Scala, 3rd Edition
Martin Odersky
55
Paperback
S$69.59 prime

Kafka - The Definitive Guide
Neha Narkhede
19
Paperback
3 offers from S$40.85

Designing Data-Intensive Applications: The Big Ideas Behind Reliable,...
Martin Kleppmann
221
Paperback
S$49.30 prime

Stream Processing with Apache Spark
Francois Garillot
1
Paperback
S$42.42 prime

USE OF DATA

google.com/search?q=big+data&oq=big+data&aqs=chrome.0.69i59j0l2j69i60l3j69i65l2.1055j0j7&sourceid=chrome&ie=UTF-8

Google

big data

All | Images | News | Videos | Books | More | Settings | Tools

About 7,230,000,000 results (0.56 seconds)

Big Data | Insights Without Limits | IBM.com

Ad www.ibm.com/IT/BigData

Infrastructure Optimized to Deliver Actionable, Game-Changing Insights. Software & Frameworks. Pure Processing Power. Simplified Software. Flexes w/ Your **Data** Needs.

The Future of Analytics
IBM® Power Systems Servers Deliver a Scalable Analytics Platform.

Manage Data at Scale
Get Advanced Storage Management with IBM® Spectrum Scale.

Big Data | Data Mining & Statistics | psb-academy.edu.sg

Ad www.psb-academy.edu.sg/ 6390 9000

Conducted by Massey University and PSB Academy, in City Campus at Marina Square. Master of Analytics programme with SAS certification, and Finance/Marketing specialisation. Courses: **Data** Analytics, Business Analytics, Finance Analytics.

Admission Criteria · About Massey University · Visit Us · Promotions

SAP Database & Data Management | Simpler, Faster, More...

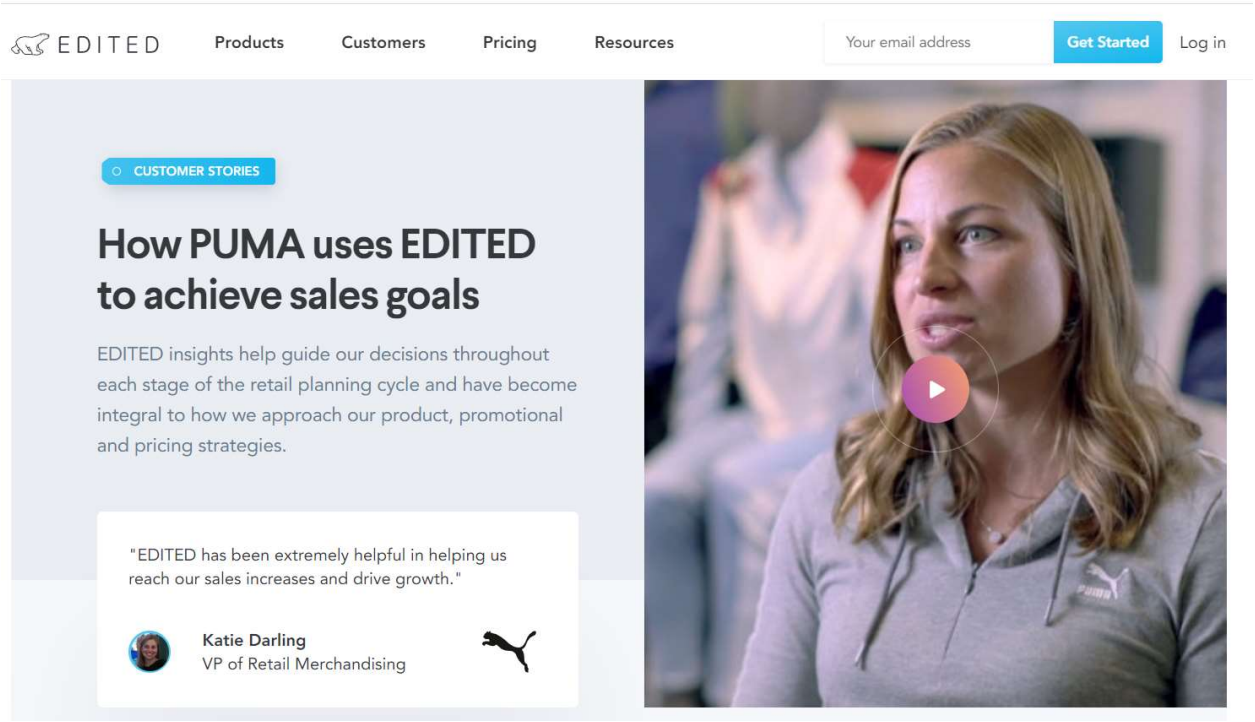
Ad www.sap.com/SEA/DataManagement

Capitalise on new opportunities w/ agile analytical & transactional technologies

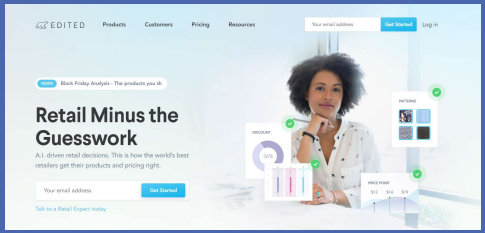
Big data

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Wikipedia

Feedback

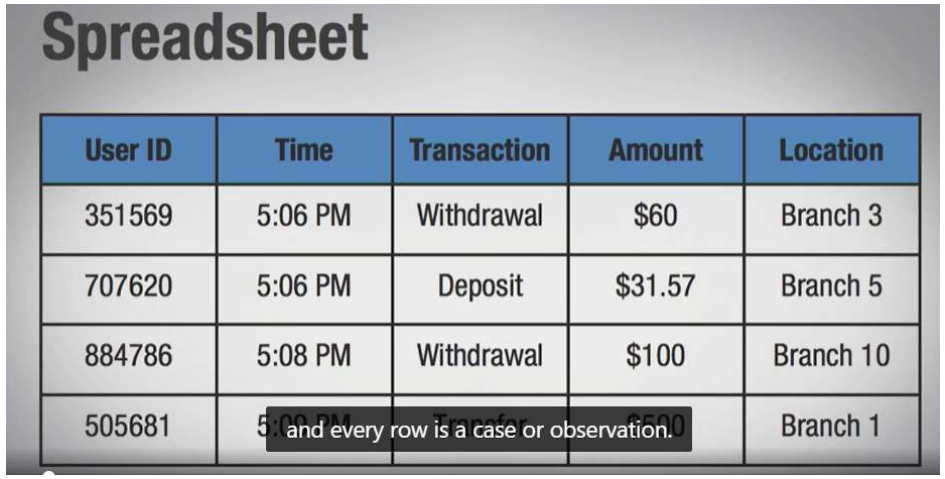


DATA FOR BUSINESS

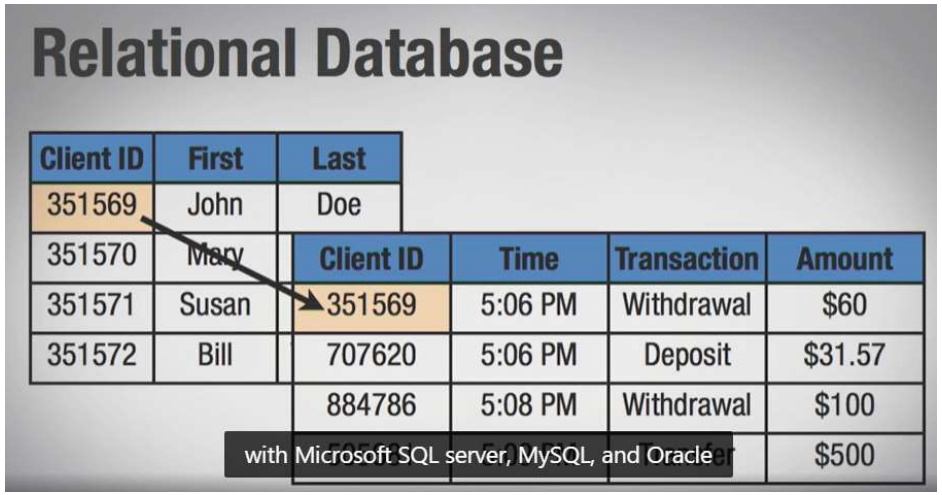


STRUCTURED, SEMI-STRUCTURED, UNSTRUCTURED DATA

- Structured Data has fixed fields.
- Easier to process by machine.



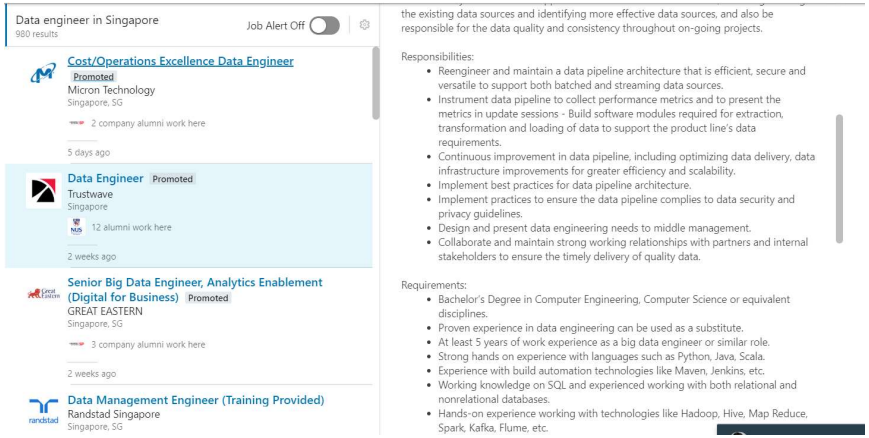
User ID	Time	Transaction	Amount	Location
351569	5:06 PM	Withdrawal	$60	Branch 3
707620	5:06 PM	Deposit	$31.57	Branch 5
884786	5:08 PM	Withdrawal	$100	Branch 10
505681	5:09 PM	Transfer	$500	Branch 1



Client ID	First	Last
351569	John	Doe
351570	Mary	
351571	Susan	
351572	Bill	

Client ID	Time	Transaction	Amount
351569	5:06 PM	Withdrawal	$60
707620	5:06 PM	Deposit	$31.57
884786	5:08 PM	Withdrawal	$100
[illegible]	[illegible]	[illegible]	$500

STRUCTURED, SEMI-STRUCTURED, UNSTRUCTURED DATA



- Unstructured Data does not have fixed fields
- Text, images, presentations
- More challenging for machine to process

STRUCTURED, SEMI-STRUCTURED, UNSTRUCTURED DATA

- Semi-structured data does not have fixed fields
- But fields are marked, data are identifiable.
- Two popular formats, XML, JSON.

This XML file does not appear to have any style information associated with it. document tree is shown below.

```
▼<CATALOG>
 ▼<PLANT>
    <COMMON>Bloodroot</COMMON>
    <BOTANICAL>Sanguinaria canadensis</BOTANICAL>
    <ZONE>4</ZONE>
    <LIGHT>Mostly Shady</LIGHT>
    <PRICE>$2.44</PRICE>
    <AVAILABILITY>031599</AVAILABILITY>
  </PLANT>
 ▼<PLANT>
    <COMMON>Columbine</COMMON>
    <BOTANICAL>Aquilegia canadensis</BOTANICAL>
    <ZONE>3</ZONE>
    <LIGHT>Mostly Shady</LIGHT>
    <PRICE>$9.37</PRICE>
    <AVAILABILITY>030699</AVAILABILITY>
  </PLANT>
 ▼<PLANT>
    <COMMON>Marsh Marigold</COMMON>
    <BOTANICAL>Caltha palustris</BOTANICAL>
    <ZONE>4</ZONE>
    <LIGHT>Mostly Sunny</LIGHT>
    <PRICE>$6.81</PRICE>
    <AVAILABILITY>051799</AVAILABILITY>
  </PLANT>
 ▼<PLANT>
    <COMMON>Cowslip</COMMON>
    <BOTANICAL>Caltha palustris</BOTANICAL>
    <ZONE>4</ZONE>
    <LIGHT>Mostly Shady</LIGHT>
    <PRICE>$9.90</PRICE>
    <AVAILABILITY>030699</AVAILABILITY>
  </PLANT>
 ▼<PLANT>
    <COMMON>Dutchman's-Breeches</COMMON>
    <BOTANICAL>Dicentra cucullaria</BOTANICAL>
    <ZONE>3</ZONE>
```

```
var family = {
    "jason" : {
        "name" : "Jason Lengstor
        "age" : "24",
        "gender" : "male"
    },
    "kyle" : {
        "name" : "Kyle Lengstorf
        "age" : "21",
        "gender" : "male"
    }
}
```

QUICK QUIZ

1. Online purchase will produce _____ data.
2. A posting on Facebook will produce _____ data.
3. An online recommendation system will need _____ data.
4. _____ data is easier to be processed by computer.
5. We can define the schema (structure) of _____ data, before we use the data.

6. Describe the role of Data Engineer.

THE END