Maltem Consulting Group

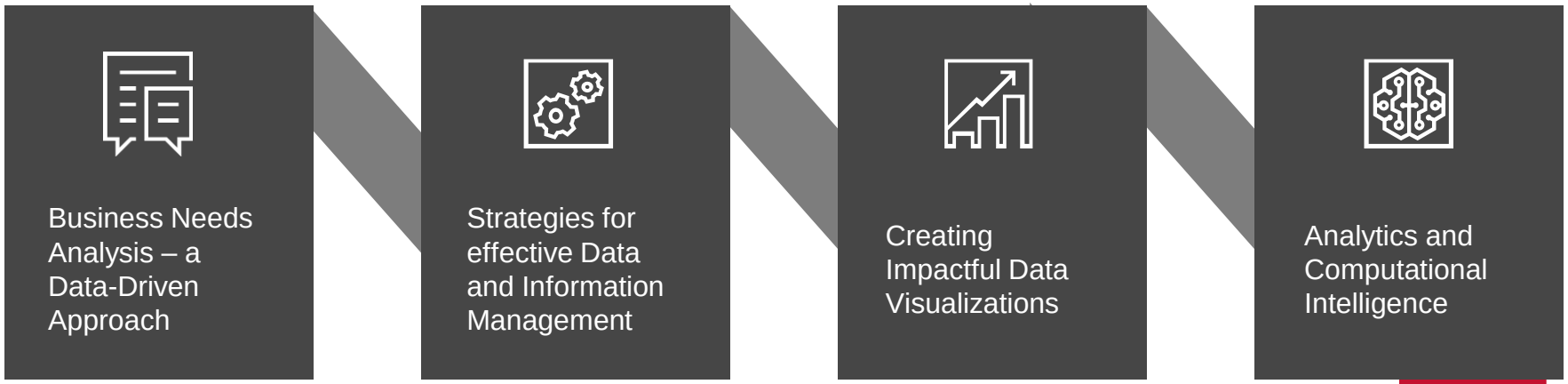
Data Engineering and Big Data

Introduction

Welcome to Data Engineering and Big Data

Data Engineers are the critical member of any Data Analytics team who are responsible for managing, overseeing, optimizing, controlling and monitoring data storage, retrieval, processing and distribution across entire organization to enable business to achieve their aspiration..

Have you taken the any courses like below?



Course outline – Day 1

Maltem Consulting Group

Background on data engineering

- What is it?
- Why do it?
- Use cases

The process flow of data engineering

- Business Requirements
- Design
- Development
- Testing
- Deployment
- Support & Maintenance

Overview of Data Engineering components

Databases

- SQL database
- NoSQL database

Data warehouse

- Data warehouse architecture
- Types of data warehouses
- Data warehouse best practices

Introduction to Big Data

- Data lakes
- Difference b/n data warehouses & data lakes

Course outline – Day 2

Maltem Consulting Group

Data Modelling

- Key Concept of Relationships
- Introduction to data modelling
- Transaction data modelling
- Analytical data modelling
- Normalization and its types
- Denormalization
- Different types of modelling techniques
- Star schema
- Snowflake schema
- Facts
- Dimension
- SCD

Data Integration

- What is data integration
- Importance of data integration
- List of data integration tools
- ETL vs ELT

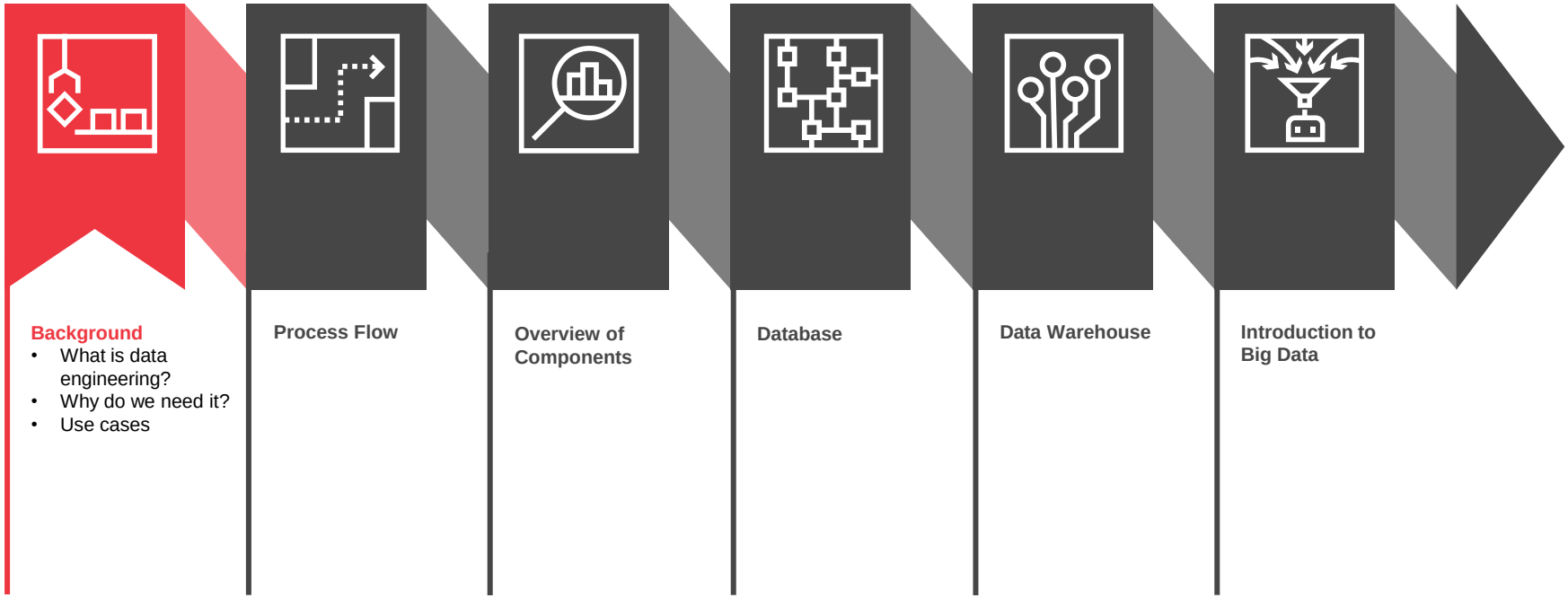
Data Integration details

- Introduction
- Operation
- Filter, Aggregation, Sort
- Join
- Union
- Business Logic

Big Data

- Hive
- Kafka
- Spark and demo

Day 1



Background
- What is data engineering?
- Why do we need it?
- Use cases

Process Flow

Overview of Components

Database

Data Warehouse

Introduction to Big Data

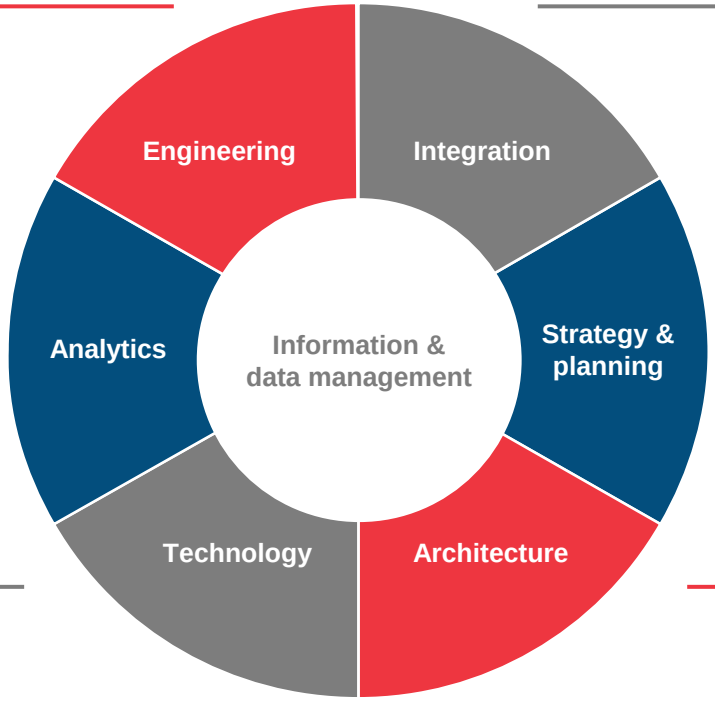
Information and Data Management disciplines

Policies, standards, SLAs, and metrics

Engineering
- Database administration
- Database programming
- Data transformation

Analytics
- Data mining
- Data analysis
- Business intelligence

Technology
- Database technologies
- Privacy and security
- Metadata management



Integration
- Data integration
- Data accessibility
- Knowledge sharing

Strategy & planning
- Data quality
- Data governance
- Data migration

Architecture
- Data warehousing
- Data modeling
- Data requirements

Data Management roles

Maltem Consulting Group

What do they do?

Data Engineer	Data Analyst	Business Analyst	Data Governance
Designs, develops and maintains data pipeline architecture and **data flow** ensuring **integrity** and **quality**. Assembles large, complex data sets from disparate system which meets functional / non-functional business requirements. 	Acquires and collects data. Develops and implements data analyses using **statistical techniques to find trends and patterns** in data sets to improve the business. 	**Identify problems and opportunities** within a organization and provide solutions that help achieve the business goals. 	**Responsible for defining and enforcing rules and policies** of data quality and security, data governance, risk management and regulatory compliance. 

Data Engineering vs Data Scientist

Maltem Consulting Group

Picture that, in a car race, the driver gets the thrill of speeding down a track, and may enjoy victory in front of a crowd, but the engineer gets the pride of fine tuning the engine, experimenting with distinct exhaust setups, and creating a powerful, robust machine.

Likewise, a data scientist can only deliver results as good as the data which he/she has access to. Most organizations store their data in a variety of formats across databases and text files. This is where data engineers come in, **building data pipelines that transform data into models and formats that data scientists can use to get better insight to drive the business**

In a nutshell, data engineering is an integral part of data science that focuses on real-world applications of data acquisition, modelling and transformations

Normally, we categorize data engineers into three groups

Generalist: responsible for every step of the data process, from managing data to analysing it.

Pipeline-centric: work alongside with data scientists to help make use of the data they collect.

Database-centric: focus on transactional & analytics databases and responsible for developing schemas and data models.

Why Data Engineers are highly valued

Maltem Consulting Group

According to Glassdoor, the average income for a data engineer is in the range of **six figures per year**, depending on the required number of years of experience and familiarity with different data engineering tools and technologies.

Primary Reasons:

- **Exponential data growth**

As per Forbes, there are **2.5 quintillion bytes of data** created each day

Volume - Significant increase in data available (e.g. social network, transaction logs)

Velocity - Fast streams of data (e.g. sensor data)

Variety - Different kinds of data (e.g. text, audio, video, images)

More organisations are looking to harness the power of data but must manage the exponential growth in data available.

- **Enable better insight into the business**

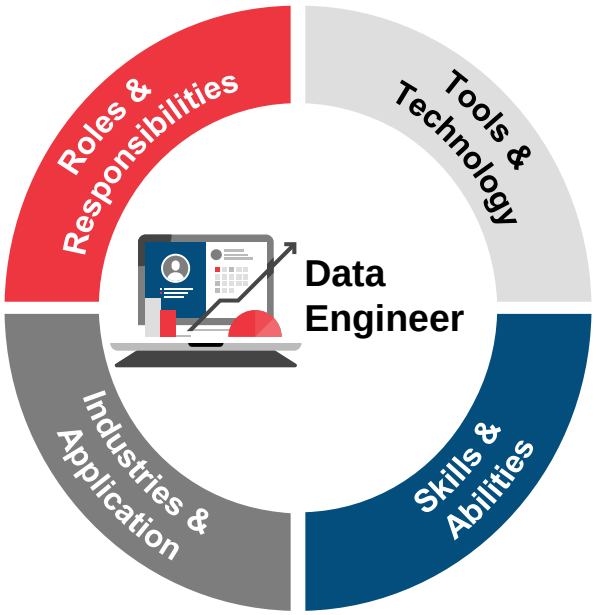
Organisations spend a lot of time, effort and money managing data to understand their product and customer needs. Data engineers can help them do so more efficiently and effectively.

- **Not enough engineers in the market**

A shortage of data engineers means that there is insufficient supply to meet the high demand for these skills. The broad and complex technical nature of the work forms a high barrier to entry that limits growth in the supply of data engineers.

What does a good Data Engineer look like?

- Designs, develops, tests and maintains data pipelines
- Builds high performance ETL solutions with industry best practices
- Expands and grows data platforms to solve and new data challenges
- Constructs and maintains data warehouse, ODS and data marts



- Relational DB - Oracle, MS SQL, Postgres etc.
- Distributed DB - MongoDB, Cassandra, Neo4j
- ETL Tools – Alteryx, Talend, SSIS, Pentaho
- Big Data – Hadoop, Yarn, Hive, Kafka, Impala, Sqoop, Spark etc.
- Coding – SQL, Python, Java, Scala, R

- Finance – fraud detection and prevention, customer segmentation, risk management
- Retail – Inventory management, recommendation, sentiment analysis
- Manufacturing – Demand forecasting, price optimization, intelligent automation

- Knowledge of data architecture
- Database management solutions
- Data warehousing and data modelling
- Designing and managing data pipelines
- Hadoop-based analytical systems

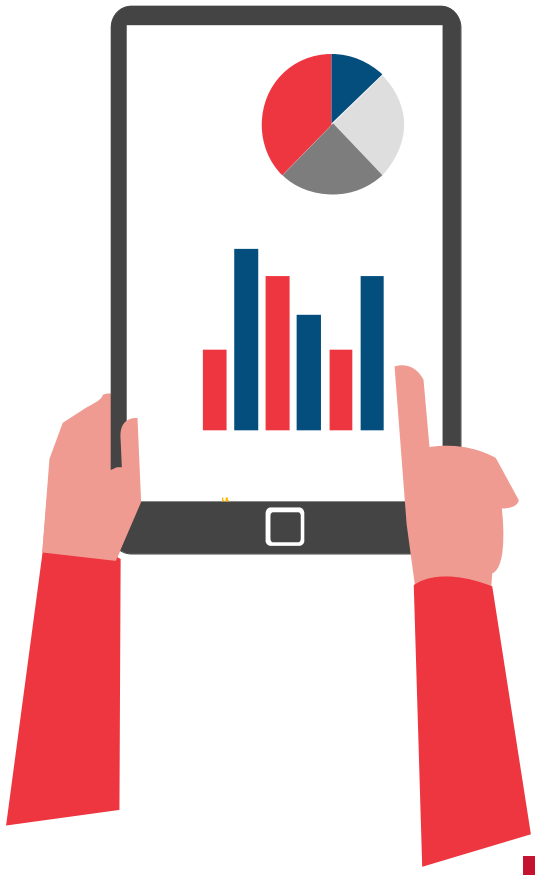
Telecom use case

Business situation:

- A telecom provider based in the US faced difficulties in sorting and integrating existing data with the data gathered by new products.
- They had to extract and maintain their data sets **manually** which was **cumbersome** and **time-consuming.**
- Moreover, with every technology update, they had to invest heavily in **updating their infrastructure**.

Solutions and outcome:

- Provided the company with an integrated view of all the relevant data within the organization with a d**ata warehousing solution.**
- With **powerful analytical dashboards**, it has become easier for the company to analyse orders, revenue statistics, maintenance contracts and more.



Retail use case (cont'd.)

Maltem Consulting Group

> The secret of successful retailing is to give your customers what they want. And really, if you think about it from your point of view as a customer, you want everything: a wide assortment of good-quality merchandise; the lowest possible prices; guaranteed satisfaction with what you buy; friendly, knowledgeable service; convenient hours; free parking; a pleasant shopping experience.
>
> **Sam Walton**
> Founder of Wal-Mart

Solutions and outcome:

Wal-Mart keeps track of 100 million customers buying billions of products every week. Using this data allows Wal-Mart to achieve "Always Low Prices". Wal-Mart can look at the sales of a given item, store by store, and determine whether something did not sell well because it was not on the floor on the best day of the week or timed with an advertising campaign.

It is the data warehouse that enabled Wal-Mart to become one of the 15 most profitable companies in the world.

Retail use case (cont'd.)

Maltem Consulting Group

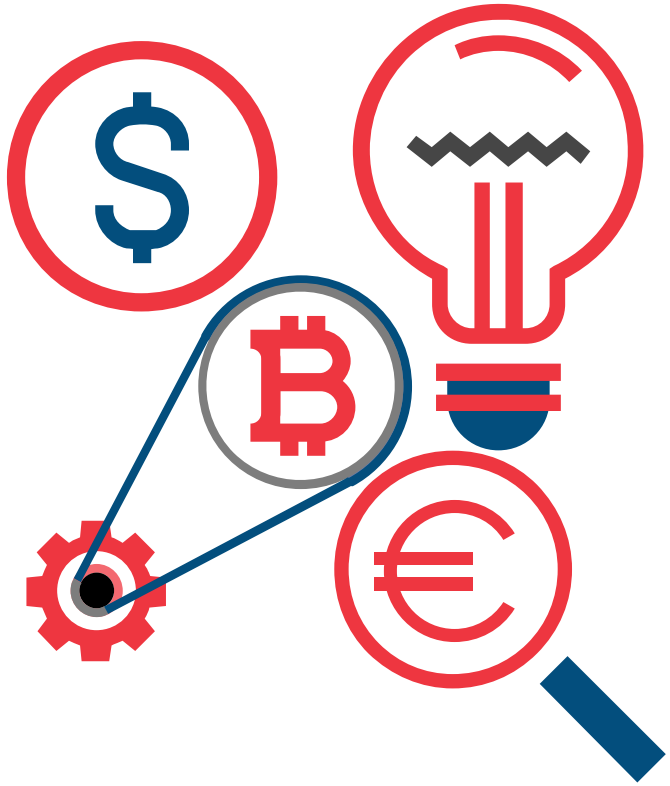
Business situation:

Over the past decade, eBay has moved well beyond its origins as a consumer-to-consumer auction website to establish its value as a mainstream platform for business-to-consumer retail and, lately, as a strategic channel for some of the world's top brands. In doing so, eBay had to dramatically raise the sophistication of the insight into transactional data that it provides to those higher-level retailers.

Solution and outcome:

eBay has spent the past two years transforming its data analysis and reporting capabilities so that its front-line staff may help themselves to its massive data store.

The e-commerce giant stores almost 90PB of data about customer transactions and behaviours to support some $3500 of product sales a second.



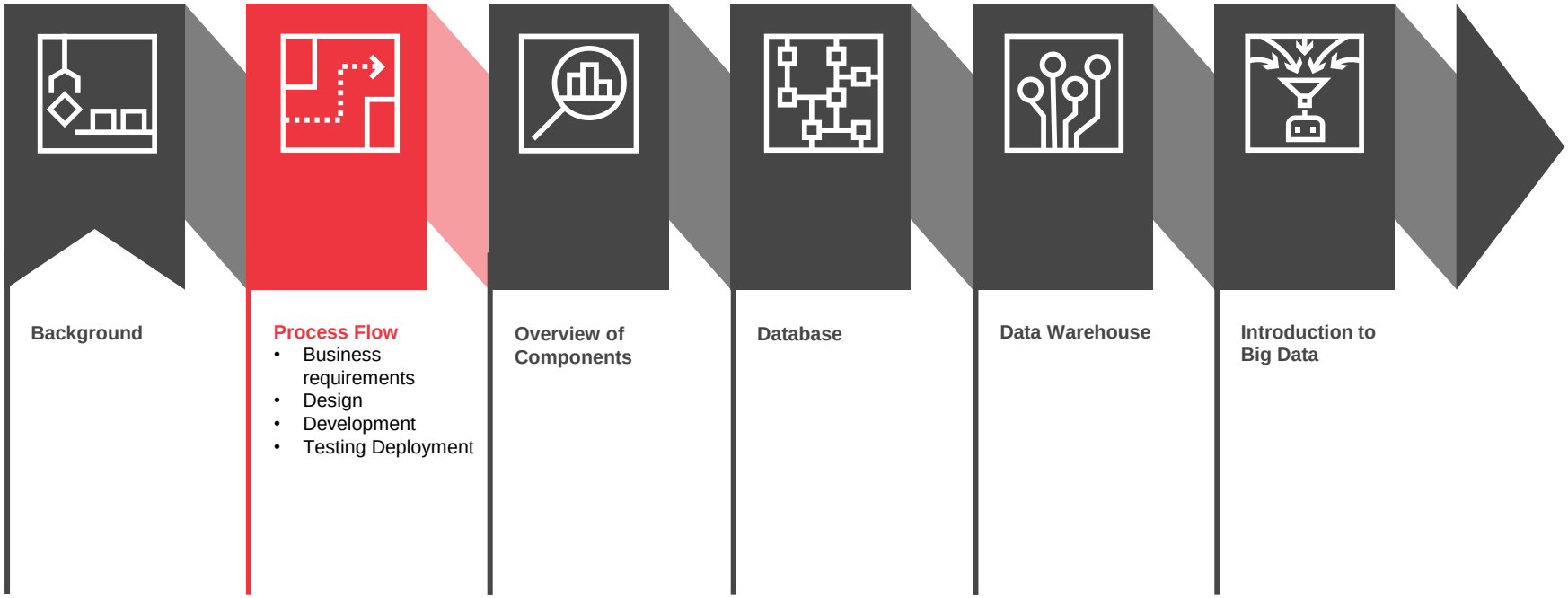
Quick Quiz

How does the work of a data engineer relate to the work of a data scientist?

What value do data engineers provide?

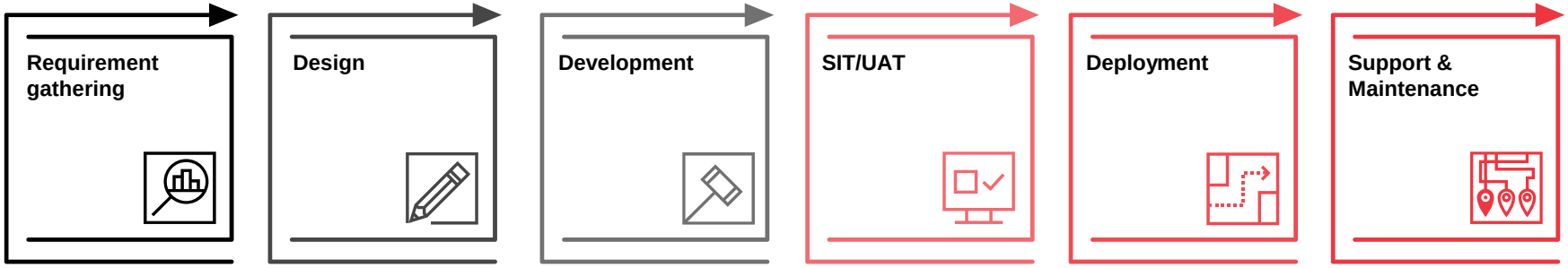
How do you become a effective data engineer?

Day 1


Background
Process Flow
• Business requirements
• Design
• Development
• Testing Deployment
Overview of Components
Database
Data Warehouse
Introduction to Big Data

Data Engineering Lifecycle

6 Phases of Data Engineer Lifecycle



Clearly define the business needs and the problems to be solved.

Understand what data will be required to address the business needs and design the data architecture involving ETL, data modelling and project timelines

Develop and implement efficient and stable processes to collect, store, extract, transform, load and integrate data at various stages in the data pipeline.

Individual components of a data pipeline are tested to ensure connections, authorization and data quality

Business users test the application which ensures that data pipeline is working as expected to address the business with predefine SLA

Operationalize the data pipeline for business applications.

Run maintenance regularly to ensure that the data pipeline works as expected. Capture feature requests for future dev.

Business Requirements

Maltem Consulting Group

Goals

- Specify the **Business Objectives**, such as increasing the sales for a retailer, or managing cost for a manufacturing company
- Identify **Key Performance Indicators (KPIs)**, such as revenue per product, that allow the business to make better decisions to achieve the business objectives
- Identify relevant **Data Sources** that the business has access to or needs to obtain, point of sales, or inventory data

How to do it?

- Identify and interact with **business stakeholders** to determine their business objectives, and understand what KPIs they use to achieve those objectives
- Identify and interact with **technology stakeholders** to determine their IT objectives, and understand what data sources, tools, and skills are available to build the solution
- Gather any relevant documents, such as system specifications and **data models** that provide details about how data is stored and manipulated to generate KPIs
- Identify business timelines, and use those to set **project timelines**

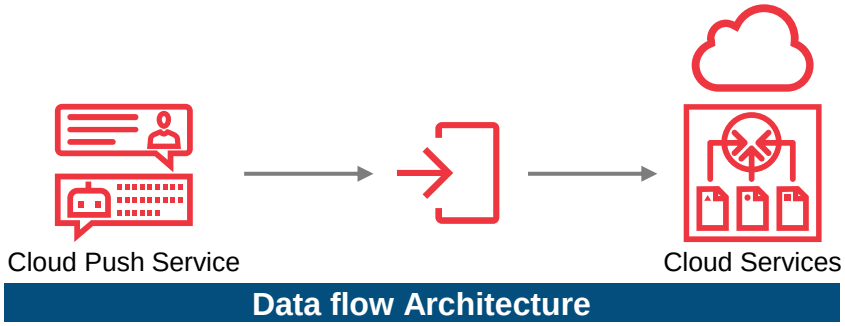
Design

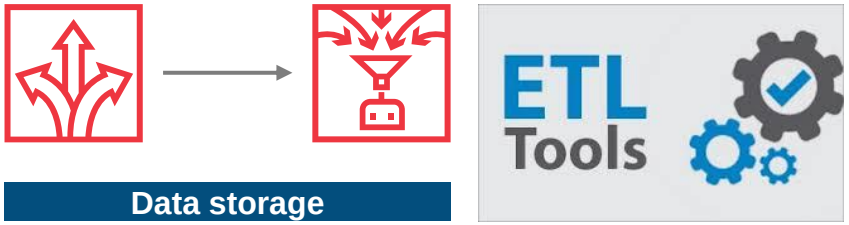
Goals

- Design the **extraction process**, determine **transformation logic aligned with the business requirements**
- Identify data integration **tools** and **relevant technologies**
- Design a **solution architecture of the data pipeline** or data workflow

How to do it?

- Start with business objectives and **work back towards data sources**
- Work with Business and IT Subject Matter Experts (SMEs) to identify required **transformation logic**
- Choose your **storage mechanism and ETL toolset** based on the volume of data and the **SLA** to provide the data to business





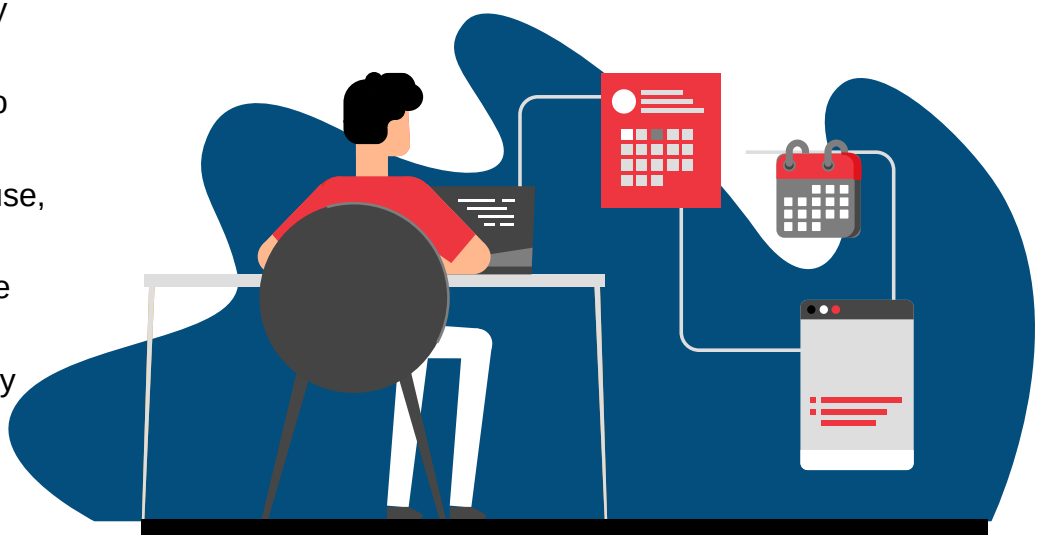
Development

Goals

- Set up a **data pipeline** or **data workflow** to collect, store, extract, transform, load and integrate data

How to do it?

- Establish **development standards** and utilise **frameworks** to accelerate development and simplify maintenance and enhancements
- **Transform** and **normalize** data from each source to match the format and schema of its destination
- **Move the data** to the target database/data warehouse, and **store** and **deliver** data in the data pipeline
- Add and delete fields and change the schema as the **business requirements change**
- **Maintain and improve** the data pipeline consistently
- Apply industry best practices



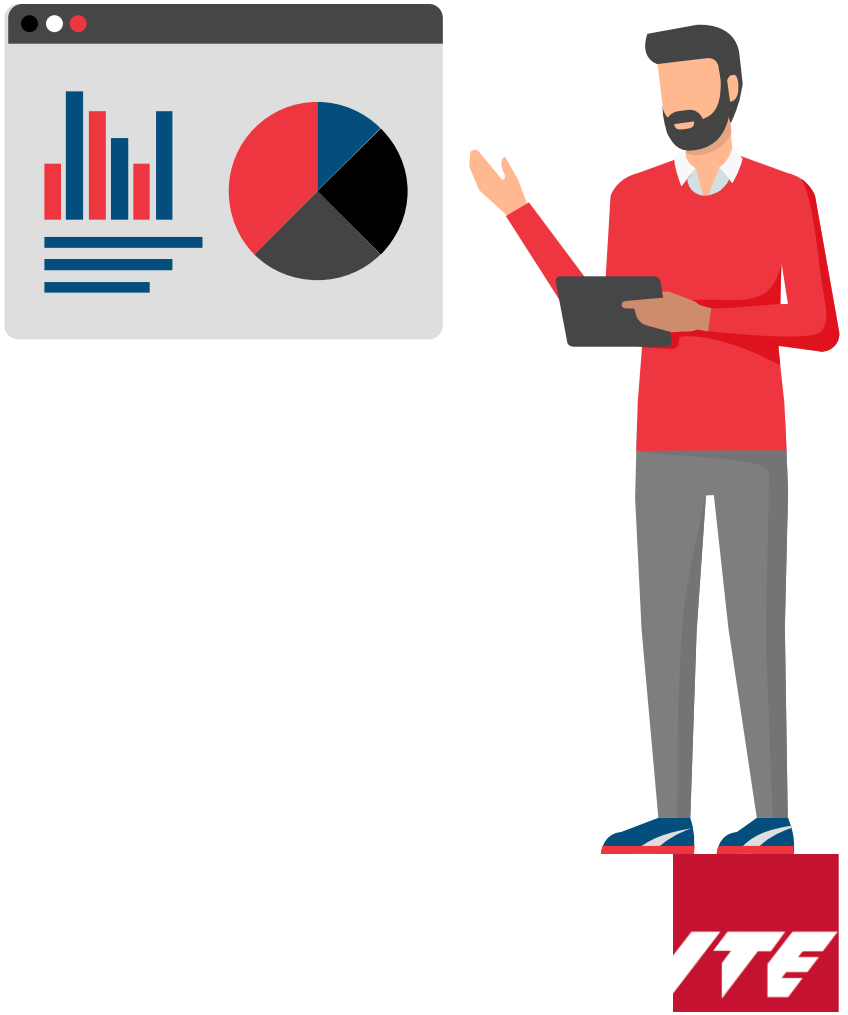
Testing (SIT/UAT)

Goals

- **Test** the **data pipeline** to ensure each component of system is **correctly interconnected** across different system and the data pipeline **built meets the business requirements.**

How to do it?

- Design test cases (SIT/UAT) and prepare test data
- **Validate** the **data required** and the **data source**
- Performance acceptance criteria (SLA)
- Validate source to target mapping
- Validation of **data models**
- Indexing, partitioning
- Error logging/Exception handling/recoverability
- ETL logic (full/incremental)
- Summary report and result analysis



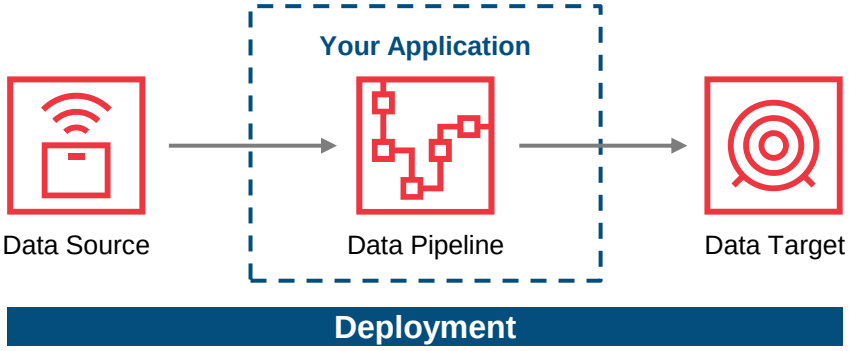
Deployment

Goals

- Deploy the model and data pipeline to production for business application.

How to do it?

- **Automate** deployment processes to accelerate code migration and reduce the risk of deployment errors
- **Operationalize** tested data pipeline for **business/downstream application** for consumption
- Expose application with an standard interface like JDBC,REST APIs
- The interface enables the data pipeline to be **consumed from various applications** such as
  - Online websites
  - Data Mining
  - Self-service dashboards
  - Internal portals
  - Back-end applications



Support & Maintenance

Maltem Consulting Group

Goals

- Increase operational efficiency and business continuity with minimum downtime to application and business.

How to do it?

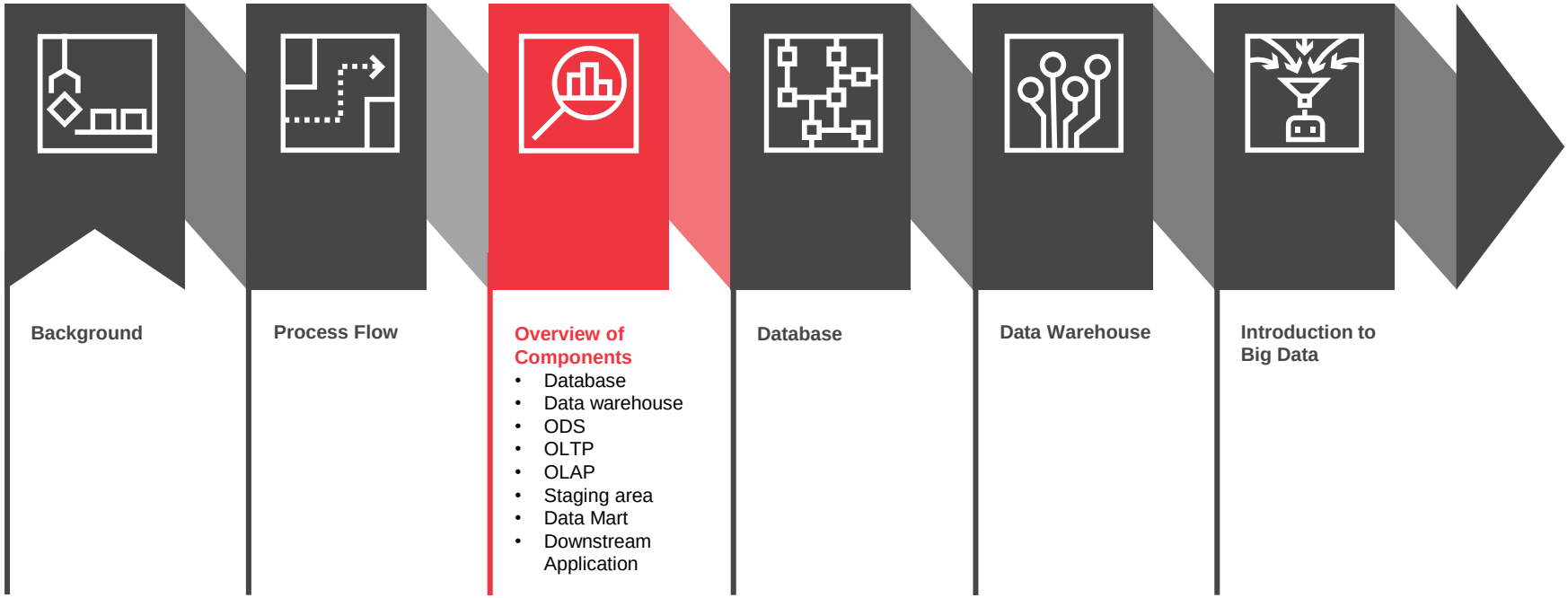
- Make sure all the 3rd party hardware and software component have **ongoing vendor support**.
- Set up a **escalation matrix** to address the critical issues and define the SLA for the resolution
- Define **change management process** to identify what changes are accepted considering cost and delivery time.
- Define the **support level** for the application depending on the criticality of the defects/issues
- Define the process to manage the **resources and their skillset** to run the project effectively

Quick Quiz

What are the different phase of Data Engineering?

What is SIT/UAT phase?

Day 1



Background

Process Flow

Overview of Components

- Database
- Data warehouse
- ODS
- OLTP
- OLAP
- Staging area
- Data Mart
- Downstream Application

Database

Data Warehouse

Introduction to Big Data

What is a database?

Maltem Consulting Group

A database is an **organized collection** of **data**, generally stored and accessed electronically from a computer system.

Wikipedia
Source

Database Management System (DBMS) - a **software application** used to create the structure to store and retrieve the data easily from the system.

- Allows data sharing
- Prevents data redundancy across application
- Increases data security
- Maintains data integrity
- Improves data consistency
- Simplifies application development
- Improves service to end users
- Enforces standards
- Provides data backup and recovery

What is a Data Warehouse?

Maltem Consulting Group

> The Data Warehouse is a database containing data from multiple operational systems that has been integrated, aggregated, and structured so that it can be used to support the analysis and decision-making process of an organization.

Ralph Kimball
Source

Comprehensive view of the Enterprise

- Data can be accessed and analysed from multiple sources
- Benefits business people to make improved and intelligent business decisions

Consistent view of the Enterprise

- Supports data conversion into a common and standard format
- The standardization of data bring consistency

Single Source of the truth

- Centralizes and integrates data from across the business enterprise
- Intuitive data models reduce IT reliance and enable business user to get faster insight

Historical Data

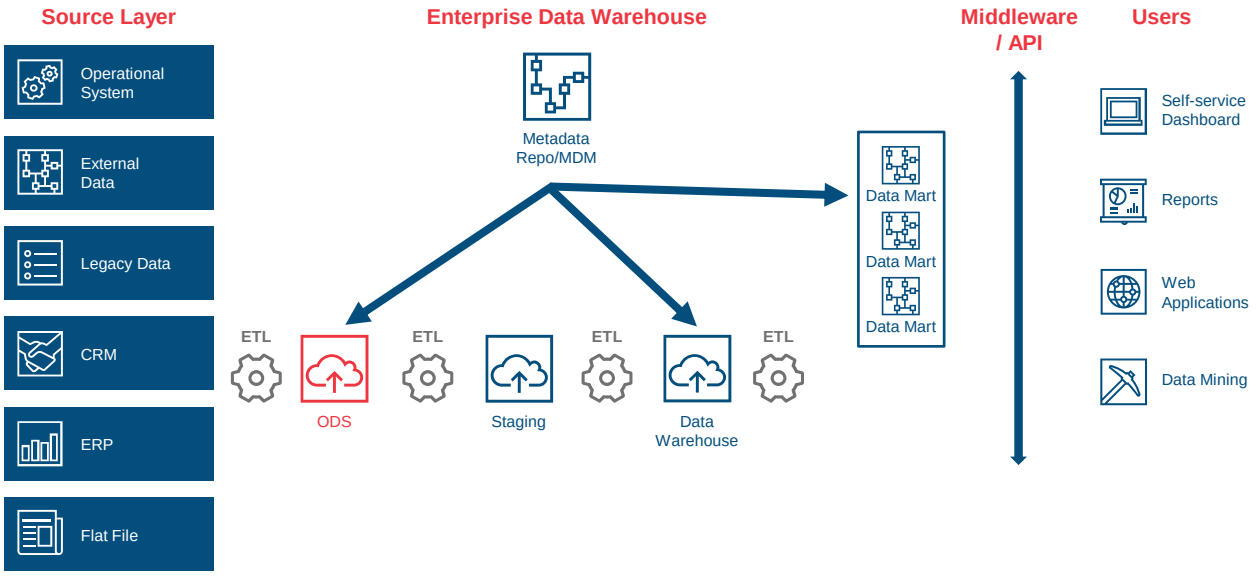
- Facilitates analysis of historical trends and patterns to predict future trends

Smarter Decisions

- Enables business users to make better decisions to improve business efficiency and drive profits

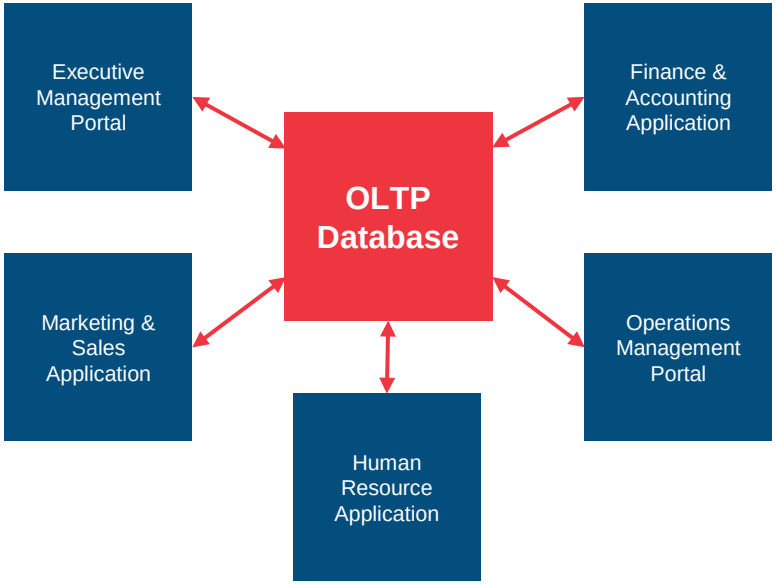
What is an Operational Data Store (ODS)?

- ODS is an **integrated database** from different heterogeneous data sources system
- Limited data persistence, **no historical data (volatile)**
- Data in ODS is not raw, during integration the data can be cleaned, denormalized, and business rules applied to ensure data integrity
- **Primarily contains current values (Active Snapshot)**
- Can be source to data warehouse and data marts
- Partial scope - contain subset of the data of an organization
- It facilitates operational reporting in real-time or near real-time



What is OLTP?

- Stands for **Online Transaction Processing**
- Objective is to efficiently process and store transactions (such as payment received from customers, products moved through inventory, orders taken from suppliers, or products delivered)
- Queries touch **small amounts of data** (one record or a few records), and **updates are frequent**
- Enables users to efficiently process business transactions and immediately make them available to clients' applications
- Maintains data integrity even with multiple application access

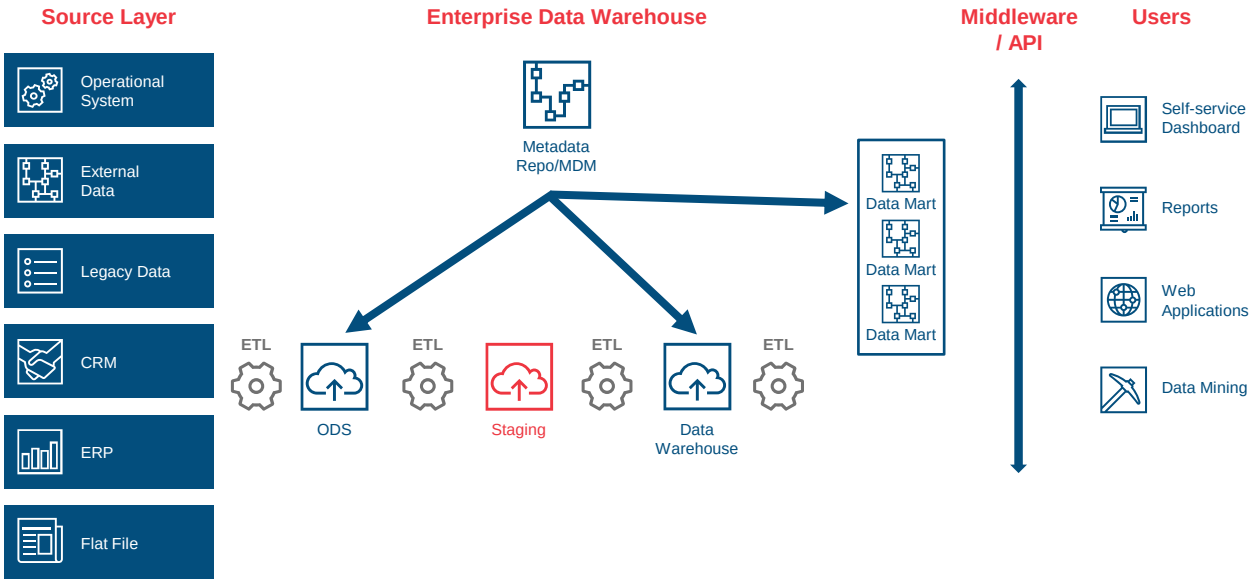


What is OLAP?

- Stands for **Online Analytical Processing**
- **Multi-Dimensional analytical database** (E.g. a company might compare their sales in June with sales in July, then compare those results with the sales from another location, which might be stored in a different database)
- Capabilities of **complex calculations**, **predictive analytics** (such as “what-if” scenarios) and sophisticated **data modelling**
- Enables users to gain **insights faster**, **interactive** and **easy to use** (BI tools) for better decision making
- The OLAP queries can include:
  - Roll Up: increasing the level of aggregation
  - Drill Down: decreasing the level of aggregation or increasing detail along one or more dimension hierarchies
  - Slice and Dice: Selection and Projection
  - Pivot: Re-orienting the multidimensional view of data

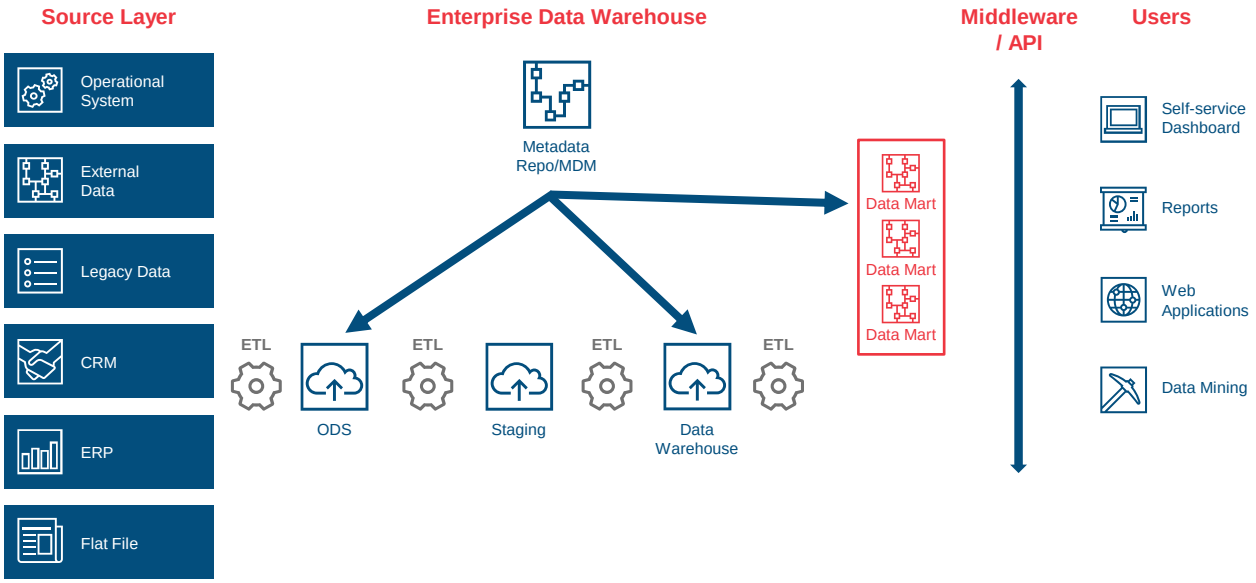
What is a Staging Area?

- The staging area is dependent on what kind of data is coming from source
- It's a **temporary place** or a phase in the architecture to hold data
- Perform **data cleansing** and **merging** before loading data in to data warehouse
- The advantage of using staging area is converting different formats of data into one format
- A disadvantage is keeping duplicate copies, but can be truncated according to business needs
- Can compare raw data with cleansed data to determine the data lineage



What is a Data Mart?

- A data mart is a **subject-oriented** database which is a partitioned segment of data warehouse
- Enables users to gain **faster access** to common data utilized
- Condensed and **more focused version** of data warehouse
- Each data mart is **dedicated** to specific **unit** or **function**
- Lower cost than implementing a full data warehouse
- Maintenance becomes **difficult** when there are **disparate and unrelated marts**



What are Downstream Applications?

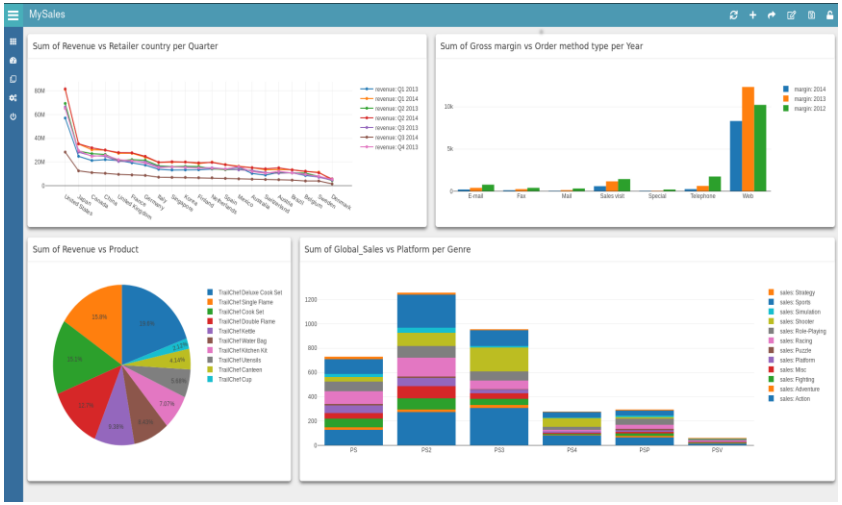
Maltem Consulting Group

Matrix showing sample use case across different industries and downstream application

	Self-service Dashboard / Reporting	Real-time Monitoring/Alerts	Machine learning / Business analysts
Manufacturing	Inventory reporting	Machinery performance tracking	Inventory management
Retailing	Sales reporting	Recommendation	Customer relationship program
Financial institutions	Credit Risk reporting	Fraud monitoring	Credit risk management

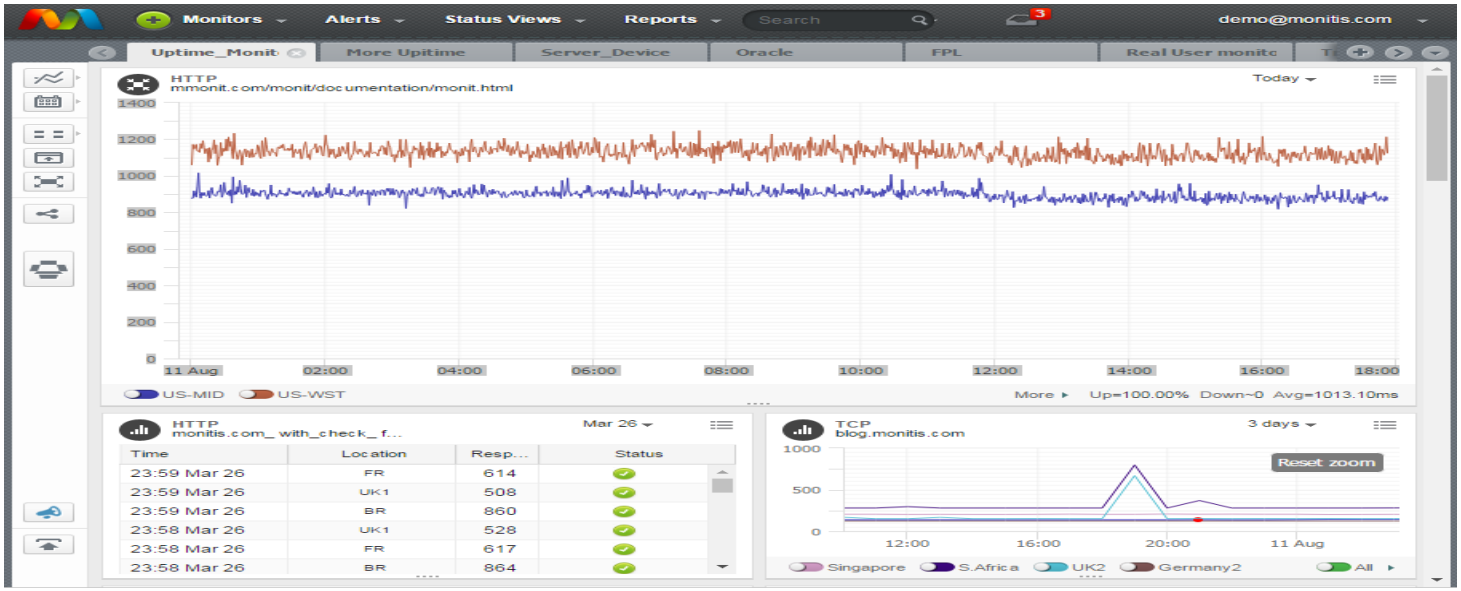
Self-Service Dashboard / Reporting

- Fast and intuitive.
- Self-service dashboard / report allow analysis based on existing data sources while **minimizing** the **query time** to pull the data
- It allows **real-time tracking** and **monitoring** various data so that business could focus more on making **data-driven decisions** and to expedite the process **of data discovery** and **insight-gathering**.
- Reduce dependency on IT.
- Increases **flexibility** to user to address key business challenges quickly and effectively.
- Requires minimal training to use
- For example, a self-service dashboard of a retail company shows the revenue, the gross margin, and sales platforms so that business users could get deep insight regarding business platform and products segments which contribute most to the sales revenue



Real-time Monitoring / Alerts

- Real Time Monitoring is a method of **observing and analysing** data as it is being accessed, manipulated, and viewed by another party
- It enables business to fix the problem right away and **stop potential damage** before it comes
- Increases productive and performance
- Proactive monitoring avoids and reduce downtime to increase business continuity
- Identify security risk and frauds.



Machine Learning / Business Analytics

- Application which has the ability to automatically learn and improve from data to make better decision and predictions
- Modified the way we extract data, interpret it by using **advance algorithms** to get the meaningful information that helps business growth
- Machine learning is used to do business analytics so that the organization could **discover value** behind the data and make **data-driven decisions**
- Machine learning helps manufacturing industry to create highly effective **predictive maintenance plan** avoiding unexpected failure and increase the productivity.
- Helps retail industry by improving the **sales accuracy** and **forecast** by combining historical selling, pricing and buying data using machine learning AI
- Help financial industry by increasing the efficiency of detecting **suspicious transaction** of **anti-money laundering** to adhere to regulatory compliance and risks
- Entertainment industry like Netflix recommending what movies and show you might like
- Autopilot/Self-driving cars

Deciding factors to choose between OLTP, OLAP

OLTP	OLAP
Its an online transaction system managing modifications	Its an online analysis and data retrieving process
Large number of short online transactions	Large volume of data
Uses traditional databases	Uses data warehouse
Normalized	Not normalized
Designed for real time business operations	Designed for analysis of business measures by category and attributes
Examples/use cases: • Online banking system • ATM • Order management system • Ticket booking system	Examples/use cases: • Data mining • Forecasting • Financial reports • Sales report

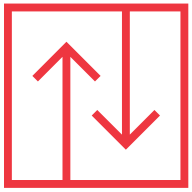
Deciding factors to choose between Data warehouse and ODS

Data Warehouse	ODS
Data Warehouse serves as a repository for a cleaned and consolidated data set	ODS serves as a channel for data between operational and analytics system
Data warehouse is updated in the batch processing	ODS is updated when transaction systems generate new data
It supports querying and reporting on the historic data	ODS is used to analyse incoming real-time incoming transaction data
Examples/ use cases: • Better understand customer and product group to know what's the buying trend • Sales reports	Examples/use case: • Operational reporting • Company owning many retail stores having its own database to track orders and we want to consolidate the databases to get real-time inventory throughout the day

Deciding factors to choose between Data warehouse and Data marts

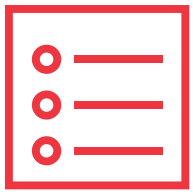
Data Warehouse	Data Marts
Data Warehouse holds multiple subject areas	It only holds one subject area, such as Finance, or Sales
It holds very detailed information	It may hold more summarized data
It works to integrate all data sources	It concentrates on integrating information from a given subject area or set of source system
Examples/use cases: • Better understand customer and product group to know what's the buying trend • Sales reports	Examples/use case: • HR data mart to address employee benefits and payroll

Quick Quiz



What is OLTP?

What are the advantages of OLAP?



Please list some downstream applications of data warehouses

Maltem
Consulting Group

End of Day 1