CM2606 Data Engineering

Introduction to Data Engineering

Week 01 | Piumi Nanayakkara

ROBERT GORDON UNIVERSITY ABERDEEN
TEF Gold
INFORMATICS INSTITUTE OF TECHNOLOGY

Learning Outcomes

- Covers LO1 for Module

- On completion of this lecture, students are expected to be able to:
  - Analyze a business process and identify how impact of big data on it.
  - Evaluate and Justify the need for Big Data Engineering for a given scenario.

CONTENT

- Impact of Data on Business
- Data Eco System
- Big Data paradigm
- Data Engineering
- Hadoop Eco System

Data Solutions in Real Life

Retail

- Demand Prediction
- Inventory Planning
- Shelf Optimization
- Personalized Content
- Personalized Marketing

Transportation

- Optimized Delivery Scheduling
- Fuel Efficiency Analysis – Airplanes
- Predictive Maintenance
- Congestion Prediction / Management

Banking & Finance

- Automatic Loan Approvals
- Predict New Branch Locations
- Fraud Detection
- Trade Surveillance

Tele-Communication

- Churn Prediction
- Up sell/ Cross Sell Products
- Geo Mapping
- New Product Development
- Price Optimizing

HealthCare

- Disease Diagnosis
- Treatment Recommendations
- Realtime Analytics via wearable devices
- Pandemics – outbreak warnings

Data use cases that changed the world…

- Amazon - Transforming E-Commerce
  - Personalized Recommendations
  - Anticipatory Shipping
  - Price Setting
- Netflix
  - Move to Big Data: from DVD Rental to a Streaming Service
  - Netflix Contest from 2006 to 2009, 1 Million USD Offer
- Facebook (and other social media) Advertising
  - Realtime tracking of user behavior
  - Movie – The Social Dilemma

Data Eco System

- Data gives the competitive advantage for an organization
- Data Eco system in an organization refer to:
  - How data is **captured, stored and processed**:
    - Sources / Tools / Infrastructure
  - How captured data is used to **make value/insight generation**:
    - Analytics
  - How the stakeholders **act upon generated insights**:
    - Application
- Eco Systems are intended to evolve over time

Data Eco System Example

Retail Organization

- **Data Capture:** E-Commerce site / social media channels / mobile apps / chat bots / call center logs /video surveillance in store
- **Storage:** Database / Datawarehouse / Data Lake
- **Process / Insight Generation:** Predict Sales / Sentiment Analysis / Queue time prediction / Staff Allocation
- **Actions:** Strategic / Operation business decision making

Data Eco System – Modern Implications

- Cloud Platforms / Cloud Data Platforms:
  - Limitless storage
  - High –performance computing,
  - Latest tools and libraries
- Machine learning
  - With availability of big data multiple avenues are created for organizations to leverage these data to improve performance
- Big Data
  - Inflow of data is changing from multiple dimensions

Big Data Characteristics

- The V's of Big Data
  - **Volume** – Enough data for the requirement
  - **Velocity** – Speed at which data comes in
  - **Variety** – heterogeneous data sources
  - **Variability** – Data from same source varies with time
  - **Veracity** – Accuracy or truthfulness of a data set
  - **Value** – Access to the data when most needed for business
  - **Visualization** – Can be processed into an understandable format
  - **Vulnerability** – Security of the source

Big Data 3Vs: Volume

There are nearly as many pieces of digital information as there are stars in the universe.

175

Zettabytes of data world wide by 2025

$ 140.9 billion

Data science platform market growth by 2024

$127 billion

AI-Based Self-Driving Car Market Is Expected to Reach

12%

Ever used and analyzed

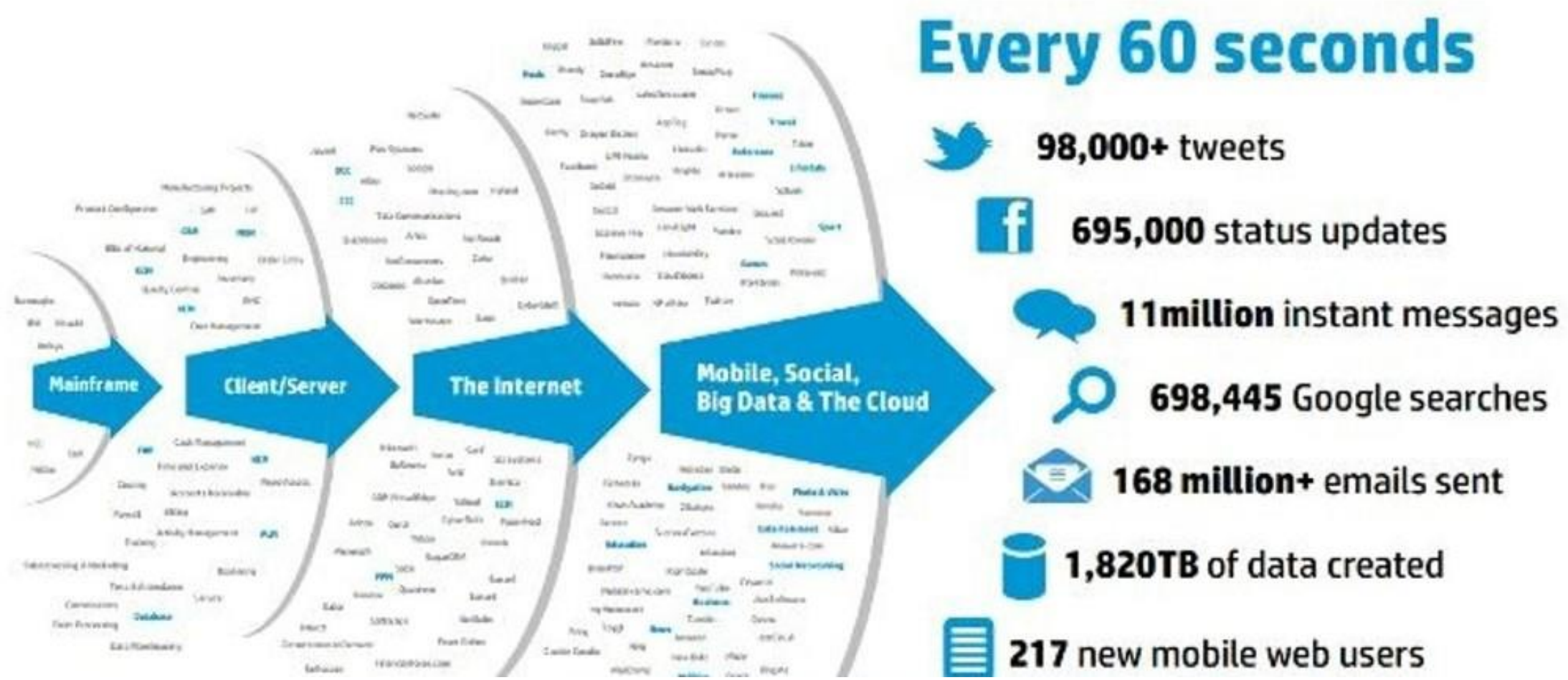
1000

Computers used by Google to search a single query

151,717

Nationally, Shortage of people with data science skills in U.S

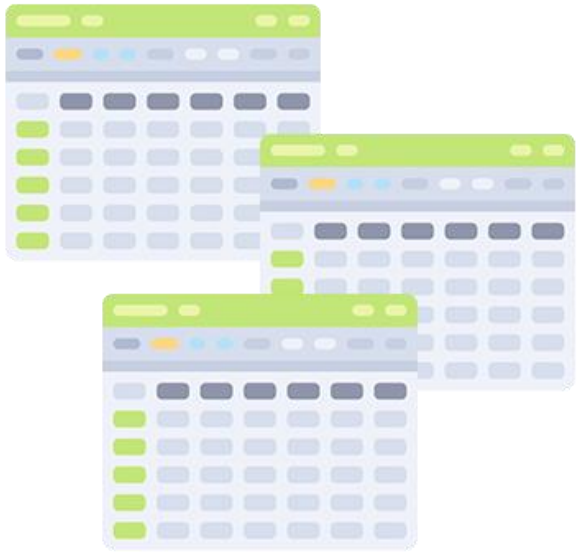
Big Data 3Vs: Velocity



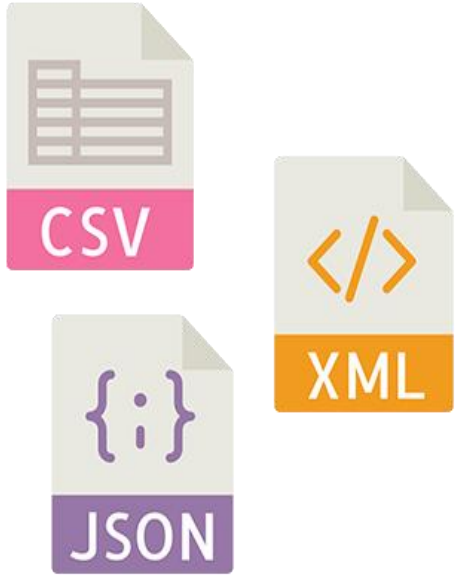
Source

Big Data 3Vs: Variety

Structured Data



Semi-Structured Data



Unstructured Data

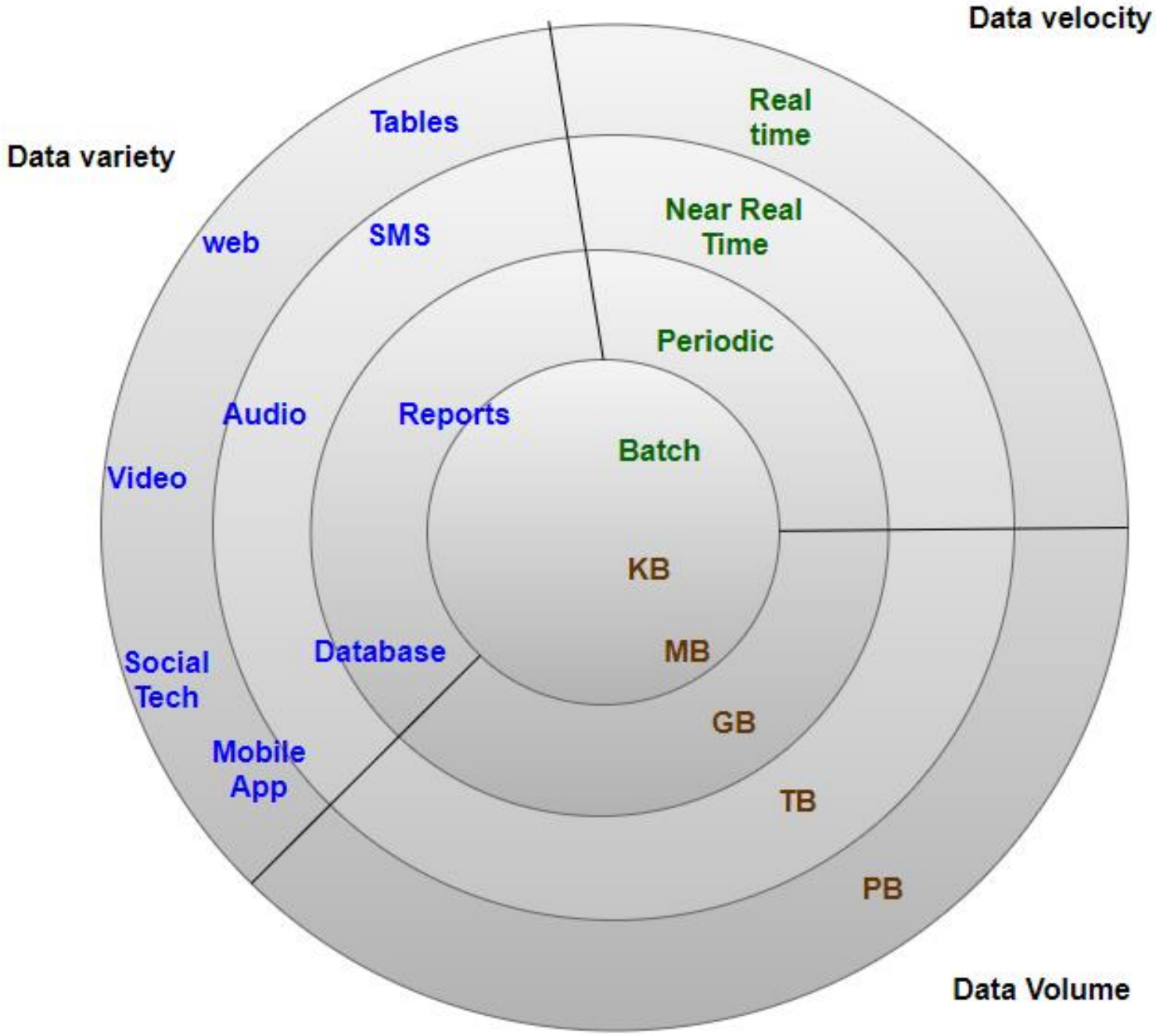
Big Data

Big Data is a phrase used to mean a

massive volume of both

structured and unstructured data

that is so large and which **moves so fast**.



Data Engineering

- Need to handle big data in production environments with required speed and accuracy.
- Responsible for making quality data available for different stake holders
- Incomes the data pipelines to serve this purpose.
  - Eliminating the manual steps and automating the process
- Software Engineers in data driven companies needed to develop tools to handle big data.

Engineering Challenges with Big Data

- How to accumulate data from multiple sources?
- How to store, move and process large volumes of data?
- How to handle unstructured data?
- How to find the insights from the huge data?
- How to filter the required data?
- How to prevent data loss?

Solution

- Data Ingestion
- Distributed Storage
- Data Warehousing
- Distributed Processing

A Data Engineer

- Around 2011 the term “Data Engineer” started to appear.

Engineers design and build things. “Data” engineers design, build and maintain pipelines that transform and transport data in a usable format to be used by entire organization.

Who is a Data Engineer

- Build and Optimize systems and pipelines handling large volumes of data. Make data accessible for anyone who needs it.

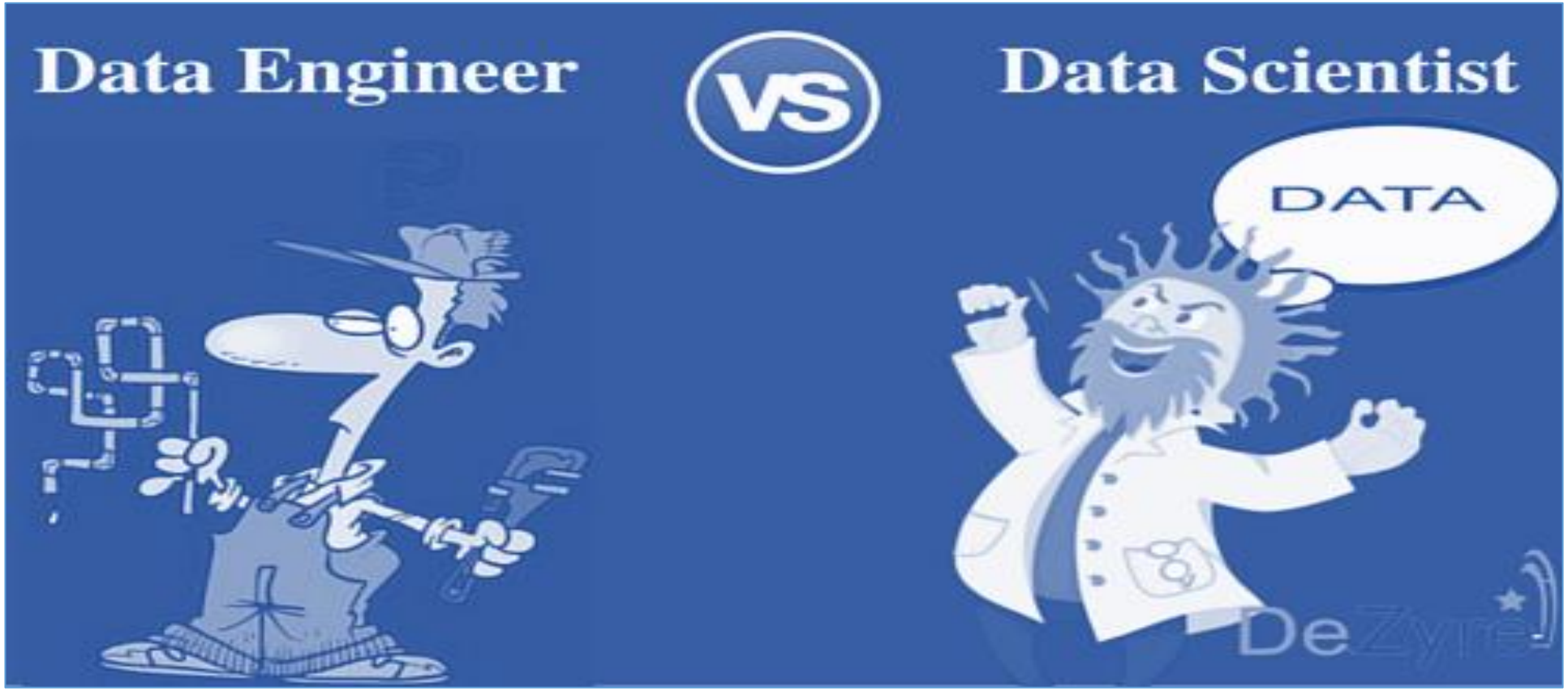
- Vs. Data Analyst
  - Analyze data and find patterns
  - Mostly Structured data

- Vs. Data Scientist
  - Deals with different types of data and different approaches to bring value out of data

- Vs. Machine Learning Engineer / Data Science Engineer
  - Take Data Science models into Production

Broader Scope…

- Role of Machine Learning Engineer / Data Science Engineer



Broader Scope…

Data Engineer

- Create and Maintain data pipelines to provide clean, quality data for data scientists

Data Scientist

- Make use of loaded data (e.g., in a data warehouse) to experiment and produce best models for the business requirement

ML Engineer

- Design and implement scheduled processes for re-training and batch scoring

Challenges of Machine Learning in Production

- An estimated 87% of data science projects never even make it to production. Why?
  - Models needs to be updated frequently
  - Data sources and types change rapidly
  - Need for Realtime / Near-Realtime Processing
  - Rollback or Failover Mechanisms
  - From Notebooks to modularized, versioned coding

- Existing Tech Stack in the Company
  - OnPrem vs Cloud Platforms
  - Tools and Techniques used
  - CICD pipelines

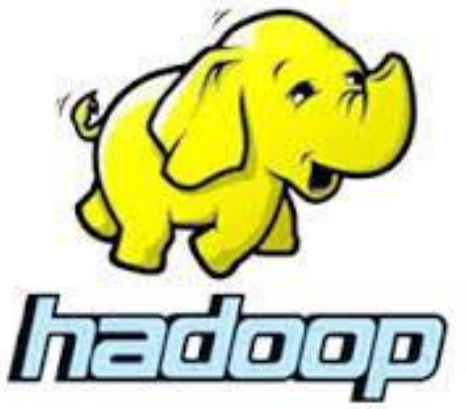
Data Engineering – Skill Set

- Programming: Python/Scala/JAVA
- Big Data Frameworks: Hadoop/Spark
- Database: SQL / Relational & Non-relational databases/ Data Modelling
- Cloud Platforms: AWS/Azure/GCP
- DevOps/Automation: Timing/dependencies/failures
- System and Technology architectures
- Machine Learning – up to some level

Hadoop Eco System

- *"An open-source software platform for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware"*

- Doug Cutting (Cloudera's chief architect) who founded the Apache Hadoop project, named after his son's toy elephant



Hadoop Eco System Components

- **HDFS** is implemented to handle large set of data. Inaugural design is brought with the inspiration of the **Google File System** (GFS).

- **YARN -** a resource management layer of Hadoop. It allows different data processing engines like graph processing, interactive processing, stream processing and batch processing to run and process data stored in HDFS.

- **MapReduce** introduced by Google. They were internally implementing ETL jobs on huge data set and they published a **Paper** that started it all. After Google's paper Amazon came up with their Hadoop instance of MapReduce is called **Elastic MapReduce** (EMR).

Hadoop Eco System: Related Projects

- **Apache Spark** an open-source alternative to MapReduce designed to make it easier to build and run fast and sophisticated applications on Hadoop It includes Spark SQL for SQL and structured data processing, **MLlib** for machine learning (ML), **GraphX** for graph processing etc.

- **Apache Pig** is a platform for analyzing huge data set on Hadoop. It's a high-level language, enables data workers to write complex data transformations without knowing Java.

Hadoop Eco System: Related Projects

- **Hive** is considered as the data warehouse. It's SQL like data analyzing framework for big data on Hadoop. This language is called **HiveQL**. Hive is mainly used for batch processing i.e., OLAP

- **Apache Tez** is generalized data-flow programming framework, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases

- **Apache HBase** is a column-oriented distributed data store. Its design and develop is inspired by Google's Bigtable.

Further Reading Material

- Hadoop: The Definitive Guide 4th Edition by Tom White, O'Reilly (2015).
- Extensive article series by Robert Chang Data Professional at Airbnb and Twitter(Former) (2018).