Data Engineering and Big Data Masters Program

Agenda

01 What is Big Data?

02 Evolution of Big Data

03 Dimensions of Big Data

04 Common Problems & Solutions(Hadoop, Spark, Cloud etc)

05 Applications of Big Data

06 QnA

What is Big Data?

According to Gartner, Big Data is a high-volume, high velocity, and high-variety information asset that demands cost-effective, innovative forms of information processing for enhanced insight and decision making

Billions of *mobile phones*
Many use tablets

B*illions of Mobile web subscribers*

B*illion content pieces shared every day* in TBs

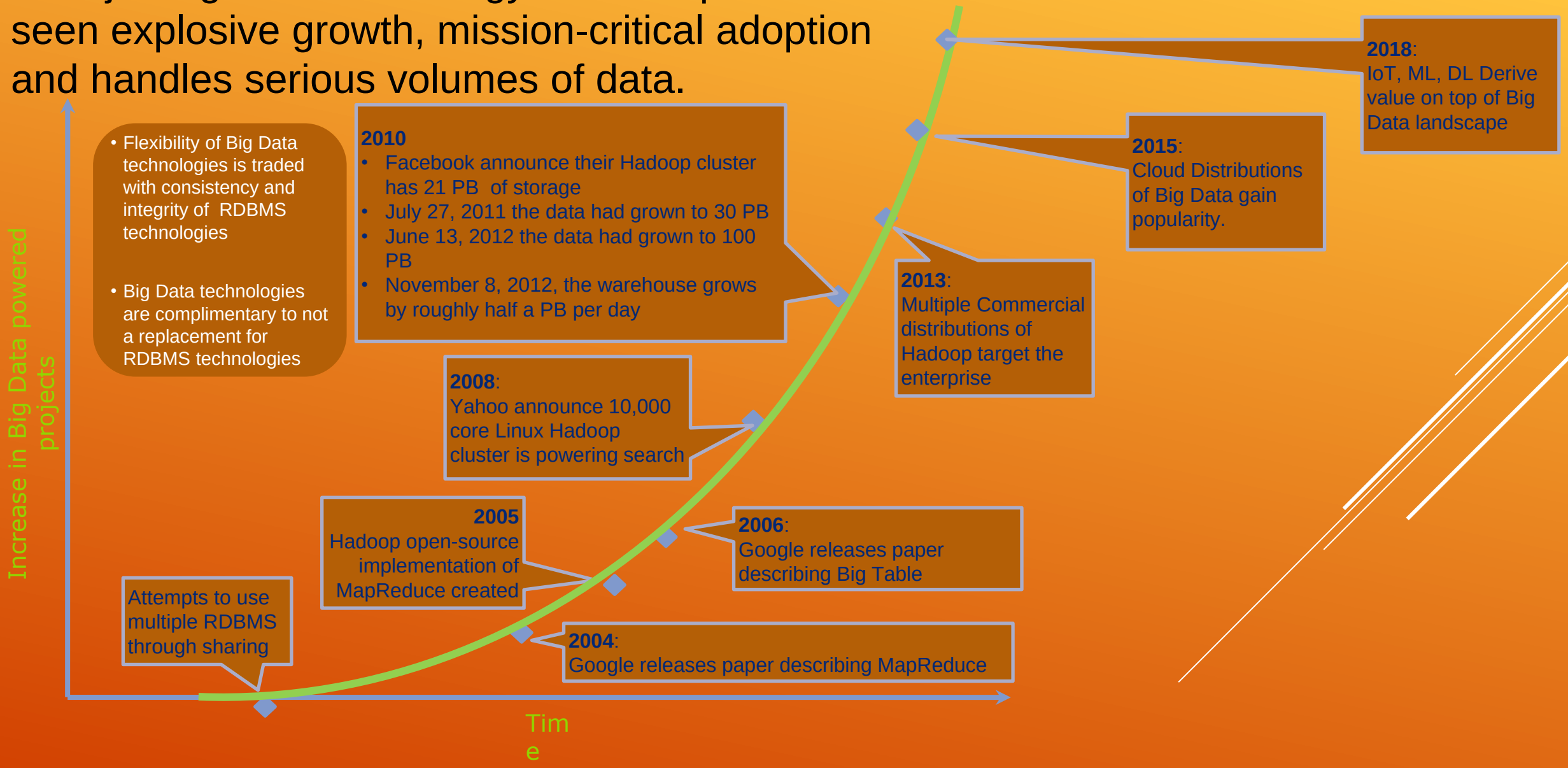
Millions of tweets *generated every day*
@Twitter 57% of all Tweets being generated in the US

B*illions of intelligent devices*

TB's *stored data in every domain*

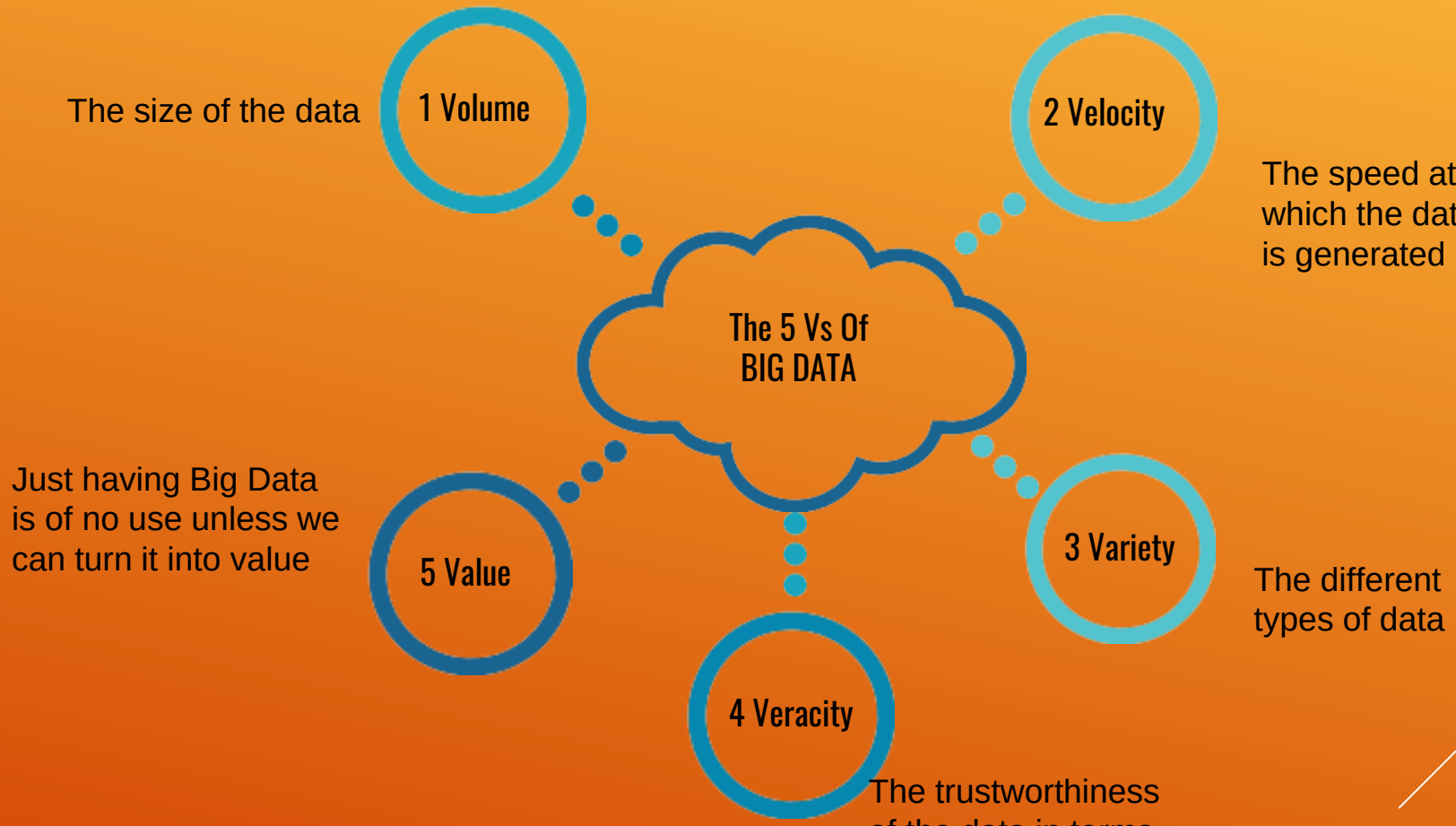
Evolution of Big Data

A major Big Data technology is Hadoop. It has seen explosive growth, mission-critical adoption and handles serious volumes of data.



Dimensions of Big Data

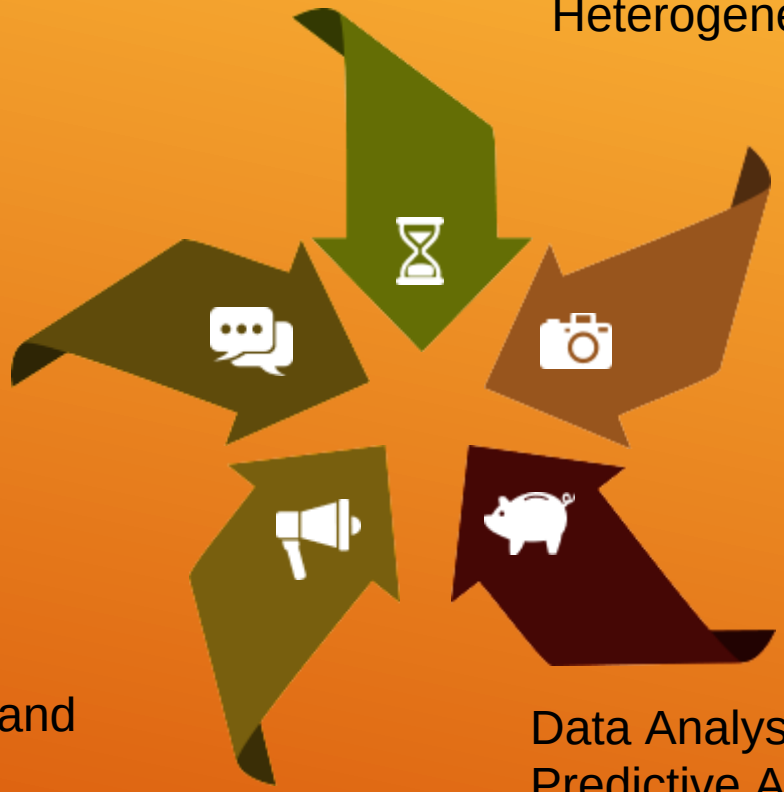
Big Data is often described by the 3 V's, each of which is a hard problem for relational databases.



Common Problems - Traditional Systems

Unimaginable size of data.

Heterogeneous systems



Traditional systems do not scale up

RDBMS is costly

Building single system is complex and not cost effective

Data Analysis, Machine Learning and Predictive Analysis do not have a common infrastructure

Possible Solutions

Scale Up

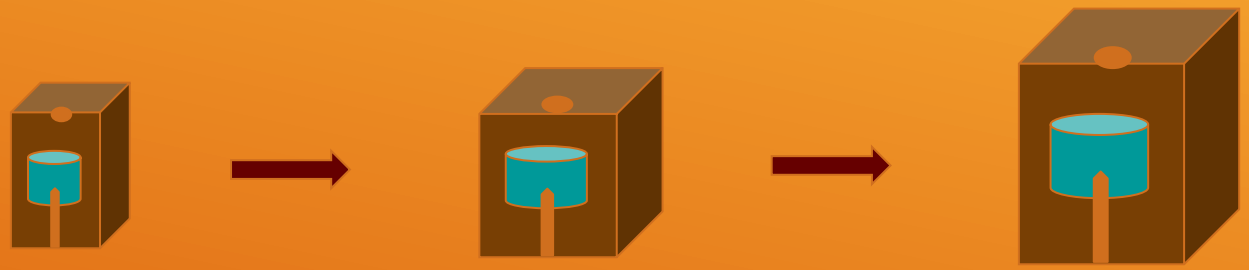
- Increase the configuration of a single system, like disk capacity, RAM, data transfer speed, etc.
- Complex, costly, and a time consuming process

Scale Out

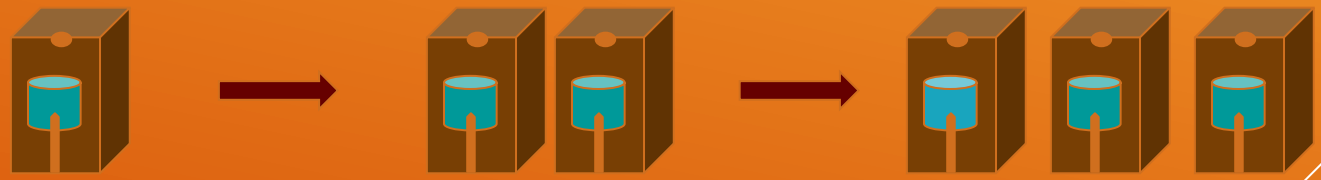
- Use multiple commodity (economical) machines and distribute the load of storage/processing among them
- Economical and quick to implement as it focuses on distribution of load
- Instead of having a single system with 10 TB of storage and 80 GB of RAM, use 40 machines with 256 GB of storage and 2 GB of RAM

Scaling Up Vs. Scaling Out

Scale-UP



Scale-Out

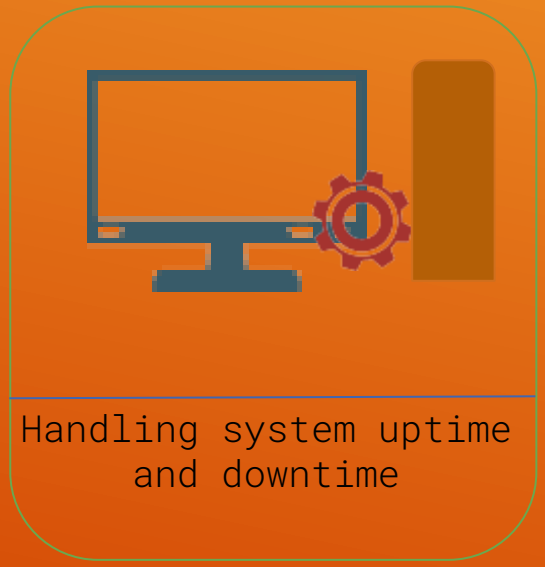


Challenges of Scaling Out

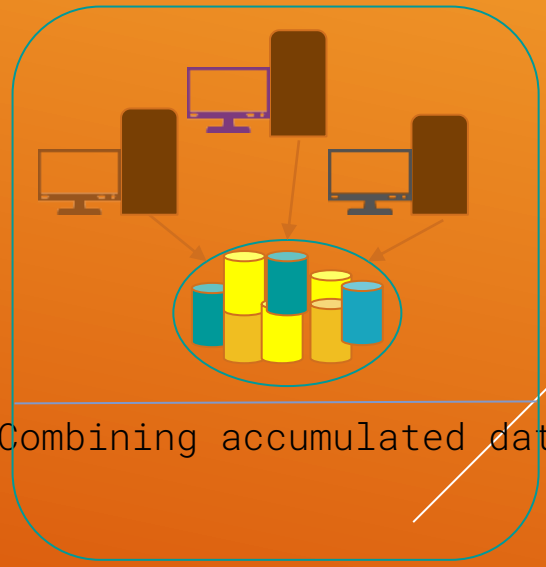
- **Need a new system:**
  - With new database management other than Relational Databases capable of handling unstructured as well as structured data
  - To process huge datasets on large clusters of computers than on a single system
- **To manage clusters in which:**
  - Nodes fail frequently
  - Number of nodes keep changing
  - Take care of communication between the nodes
  - During analysis, take results from different machines and merge/aggregate them.
- **Common infrastructure which is:**
  - Efficient
  - Easy to use
  - Reliable

Challenges of Scaling Out (Contd.)

- Big Data technology has to use commodity hardware for data storage and analysis. Furthermore, it has to maintain a copy of the same data across clusters
- Big Data technology has to analyze data across different machines and then merge the data



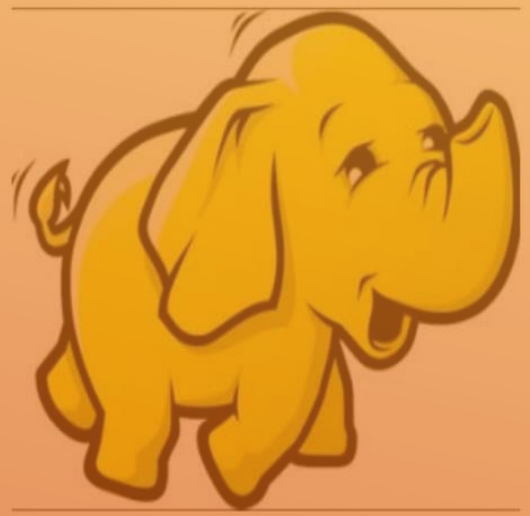
Handling system uptime and downtime



Combining accumulated data

Solution for Data Explosion - Hadoop

- Hadoop is that new framework!



What is Hadoop ?

- Open source distributed data processing cluster
- Data processed in Hadoop Distributed File System (HDFS)
- Hadoop is that new framework which helps to solve the problem of Data explosion.
- Hadoop is an open source, Java-based programming framework that supports the processing of large data sets in a distributed computing environment
- Hadoop provides: A reliable, scalable platform for storage and analysis
- It is based on Google File System or GFS
- Hadoop runs a number of applications on distributed systems with thousands of nodes involving petabytes of data
- It has a distributed file system, called the Hadoop Distributed File System or HDFS, which enables fast data transfer among the nodes
- It leverages a distributed computation framework called MapReduce
- Resource Management is performed by YARN

Hadoop: Introduction (Contd.)

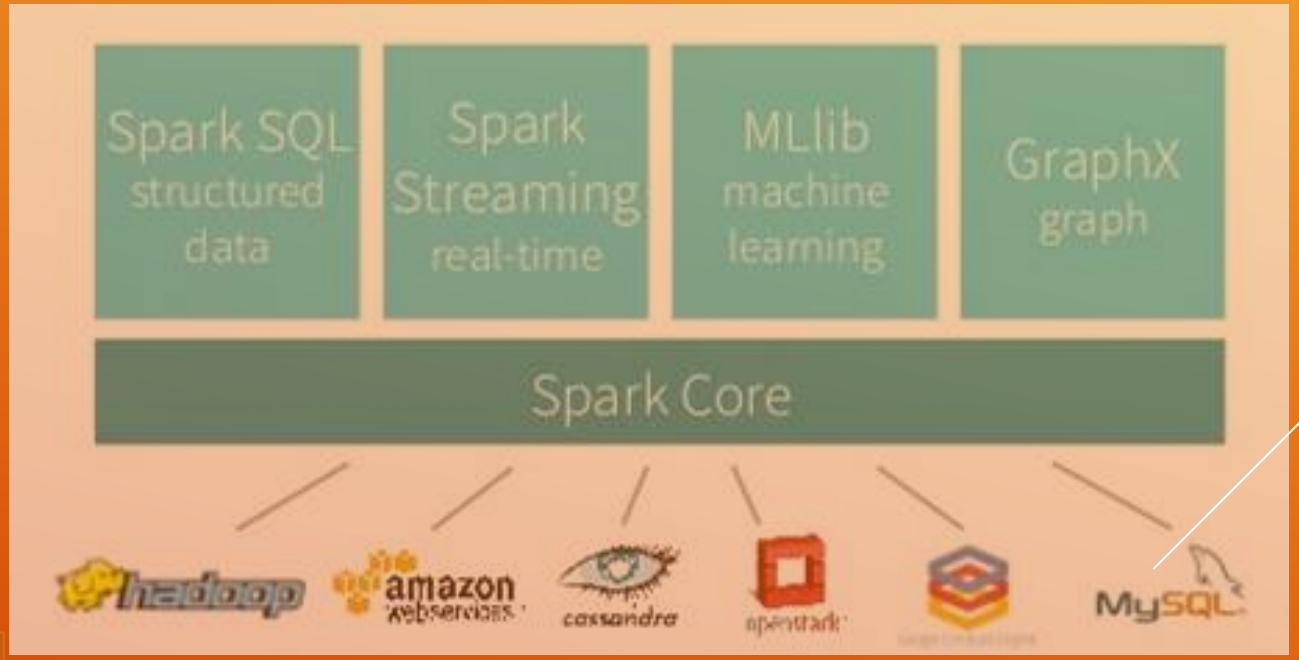
- Problems with distributed processing:
  - Hardware failure: can be solved by redundancy
  - Coordinating the tasks and combining results from all machines
- Hadoop takes care of the above complexities and the challenges of network/distributed programming
  - HDFS (for storage)
  - Map Reduce (for processing)
- Two key concepts:
  - Storage (of data and results)
  - Processing (Analysis of data)

Tools in Core Hadoop Ecosystem

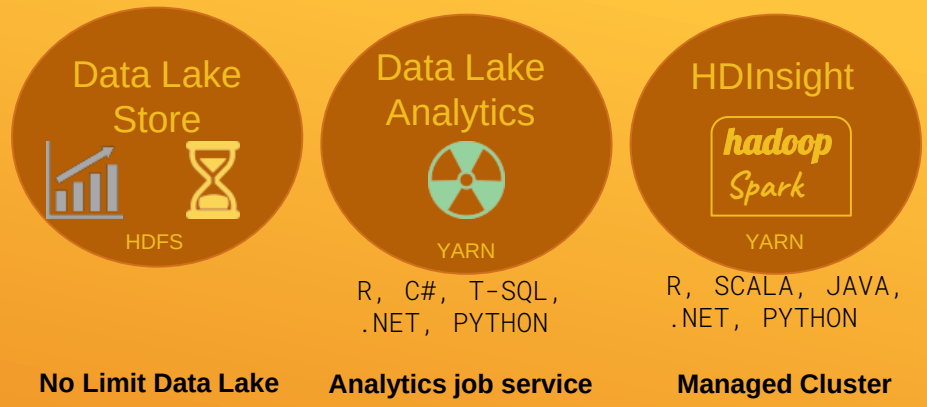
Hadoop HDFS– Distributed storage layer for Hadoop. Yarn Hadoop – Resource management layer introduced in Hadoop 2.x.
Hadoop Map-Reduce – Parallel processing layer for Hadoop.
HBase – It is a column-oriented database that runs on top of HDFS. It is a NoSQL database which does not understand the structured query. For sparse data set, it suits well.
Hive – Apache Hive is a data warehousing infrastructure based on Hadoop and it enables easy data summarization, using SQL queries.
Pig (Legacy)– It is a top-level scripting language. As we use it with Hadoop. Pig enables writing complex data processing without Java programming.
Sqoop – It is a tool design to transport huge volumes of data between Hadoop and RDBMS.
Oozie (Legacy)– It is a Java Web application uses to schedule Apache Hadoop jobs. It combines multiple jobs sequentially into one logical unit of work.
Zookeeper – A centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

Spark

- **Spark Core,RDD, DF, DS, SQL –** API's for basic data processing needs for batch layer.
- **Spark Streaming-** API for real time processing needs for speed layer of data pipeline.
- **Spark ML Lib –** API for Machine learning processing needs
- **Graph X-** API for needs of complex processing of Graph based data models with nodes and interactions between them.



Big Data Cloud Ecosystem

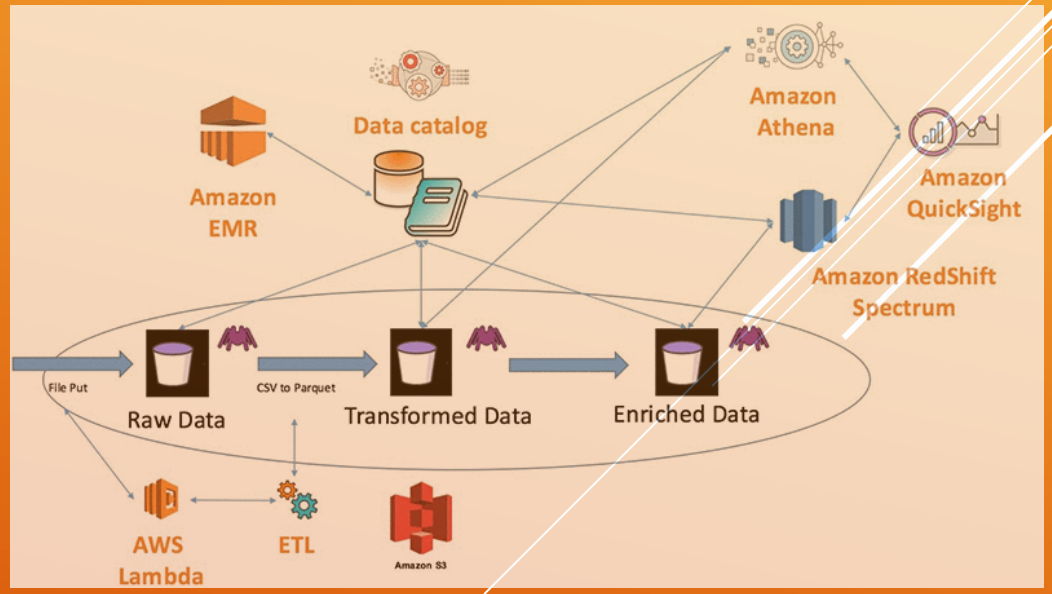


Service models based

"Infrastructure as a service" (IaaS) - VM, EC2, etc
"Platform as a service (PaaS)" - EMR, HD Insight, S3, ADLS, etc
Software as a service (SaaS) - Data Factory, Databricks, etc

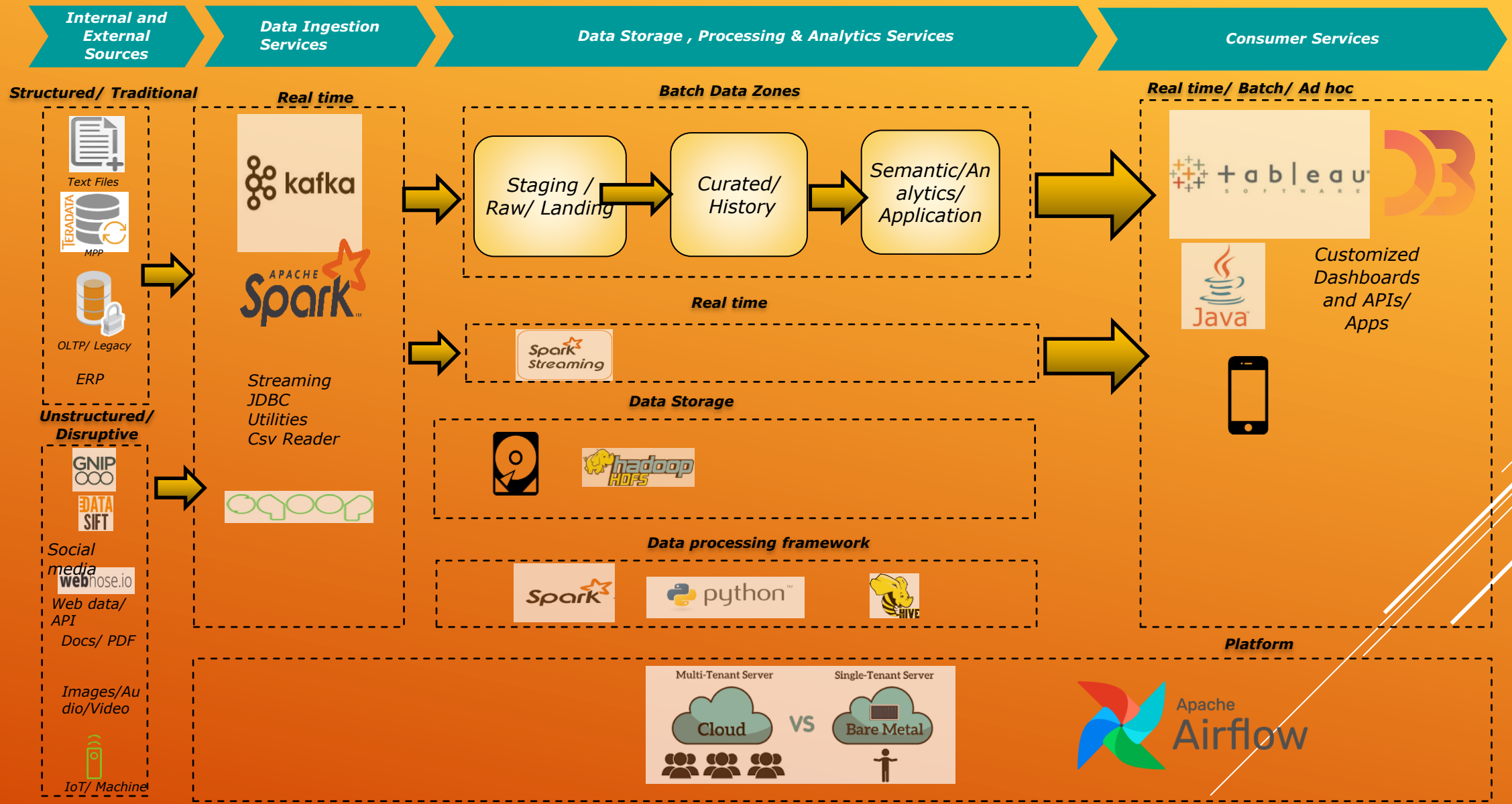
Deployment models

Private cloud Private cloud is cloud infrastructure operated solely for a single organization, whether managed internally or by a third-party, and hosted either internally or externally.
Public cloud A cloud is called a "public cloud" when the services are rendered over a network that is open for public use.
Hybrid cloud Hybrid cloud is a composition of two or more clouds (private, community or public) that remain distinct entities but are bound together, offering the benefits of multiple deployment models.
Major cloud Providers are Amazon Web Servies (AWS), Microsoft Azure Services, Google Cloud Platform, Akamai, Alibab Cloud, etc



Recommended Enterprise Big Data Architecture

iNeuron



Key Big Data Applications and their required tool/technique

Operations Analysis
Analyze a variety of machine data for improved business results

Data Warehouse Augmentation
Integrate big data and data warehouse capabilities to increase operational efficiency

Internet Of Things
Harnessing data from Sensors and actuators to make a connected world.

Key Big Data Cases in E-Commerce

Predicting Trends

Identifying Buzz from Social Media, and Ad buying data of Marketing departments

Pricing Optimization

Best pricing of a product based on competitors , etc to increase revenue

Recommendation Systems

Personalized products catalogue based on user preferences

Customer services

360 degree view of customer for Customer Support Executive and Offers

Supply Chain Optimization

- Amazon uses big data systems to select warehouses based on the proximity of vendors balanced against the proximity of customers to cut down on the distribution costs. The big data systems helps Amazon predict the number of warehouses needed and the capacity each warehouse should have.
- Links with manufacturers and tracks their inventory
- Graph theory helps decide the best delivery schedule, route and product groupings to further reduce shipping expenses.
- Predicting the products you are likely to purchase, when you may buy them and where you might need the products
- Items are sent to a local distribution center or warehouse so they will be ready for shipping once you order them
- 89% of US customers would go elsewhere for their next purchase if they face troubles with on-time order fulfilments
- Decreasing its delivery time and overall expenses and choosing the warehouse closest to the vendor and/or you, the customer, to reduce shipping costs by 10 to 40%

THANK YOU

QnA