Big Data – Session 1

Cloud Computing

Cloud Computing

- Cloud computing is the on-demand delivery of IT resources (including servers, storage, databases, networking, software, analytics, and intelligence) over the Internet with pay-as-you-go pricing

Benefits – Cloud Computing

Cost

- Eliminates the capital expense of buying hardware and software and setting up and running on-site datacenters.

Global scale

- Ability to scale elastically.

Performance

- Datacenters are regularly upgraded to the latest generation of fast and efficient computing hardware.

Security

- Many cloud providers offer a broad set of policies, technologies and controls that strengthen security, helping to protect data, apps and infrastructure from potential threats.

Speed

- Self service and on demand computing services can be provisioned in minutes, typically with just a few mouse clicks

Productivity

- Removes the need hardware setup, software patching, and other time-consuming IT management chores on resources by team.

Reliability

- Data backup, disaster recovery and business continuity is easier and less expensive because data can be mirrored at multiple redundant sites on the cloud provider's network.

Types – Based on Deployment

Public cloud

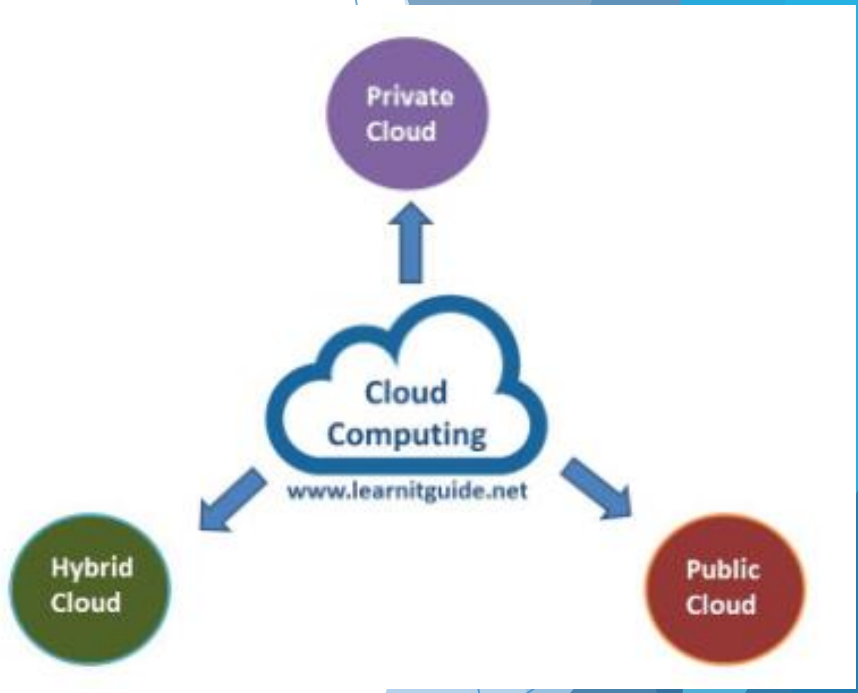
- Public clouds are owned and operated by a third-party cloud service providers, which deliver their computing resources like servers and storage over the Internet.
- Ex: Microsoft Azure, AWS

Private cloud

- A private cloud refers to cloud computing resources used exclusively by a single business or organization.
- A private cloud can be physically located on the company's on-site datacenter

Hybrid cloud

- Hybrid clouds combine public and private clouds, bound together by technology that allows data and applications to be shared between them.



Types – Based on Service Model

Infrastructure as a service (IaaS):

- With IaaS, you rent IT infrastructure—servers and virtual machines (VMs), storage, networks, operating systems—from a cloud provider on a pay-as-you-go basis
- Ex: Azure Virtual Machine

Platform as a service (PaaS):

- Platform as a service refers to cloud computing services that supply an on-demand environment for developing, testing, delivering and managing software applications.
- Ex: HDInsight, Azure Blob Storage, Azure SQL Database

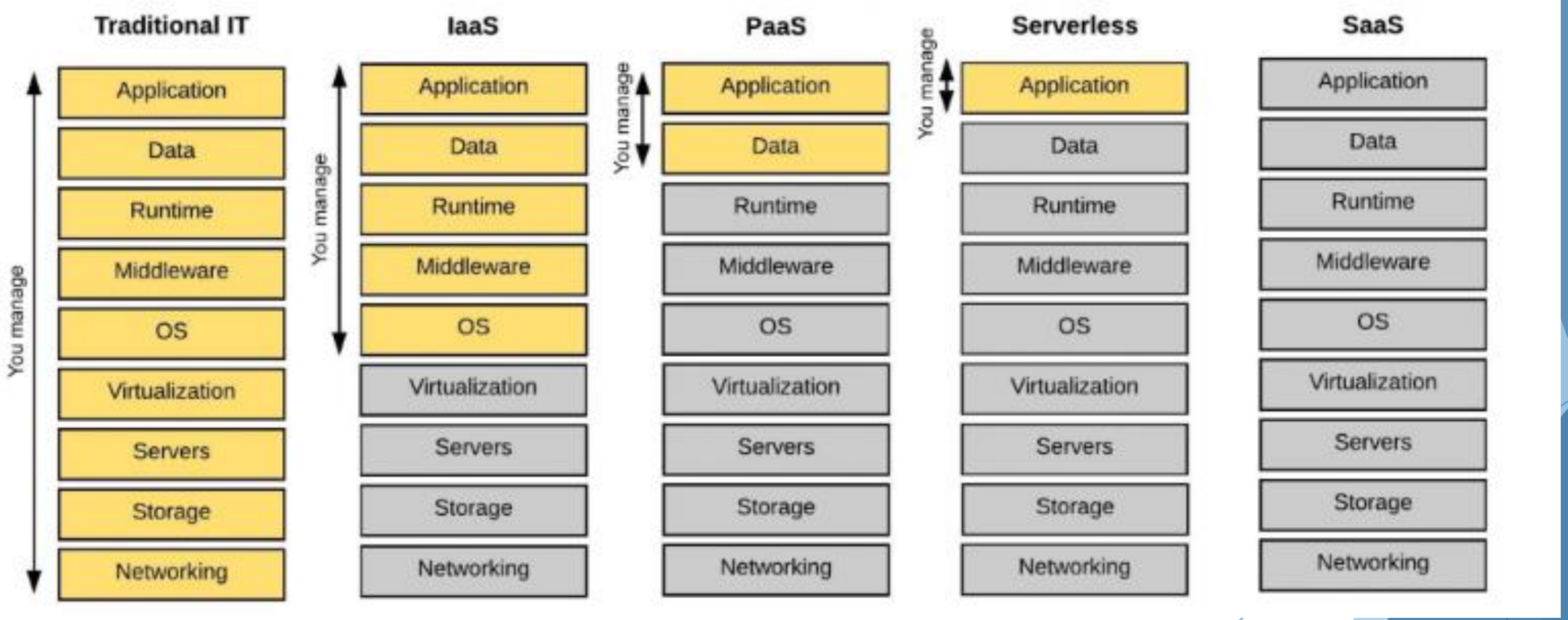
Serverless computing:

- Overlapping with PaaS, serverless computing focuses on building app functionality without spending time continually managing the servers and infrastructure required to do so.
- Azure Lamba, Azure Functions

Software as a service (SaaS):

- Software as a service is a method for delivering software applications over the Internet, on demand and typically on a subscription basis.
- Ex: Office 365, Google apps (Gmail)

Types

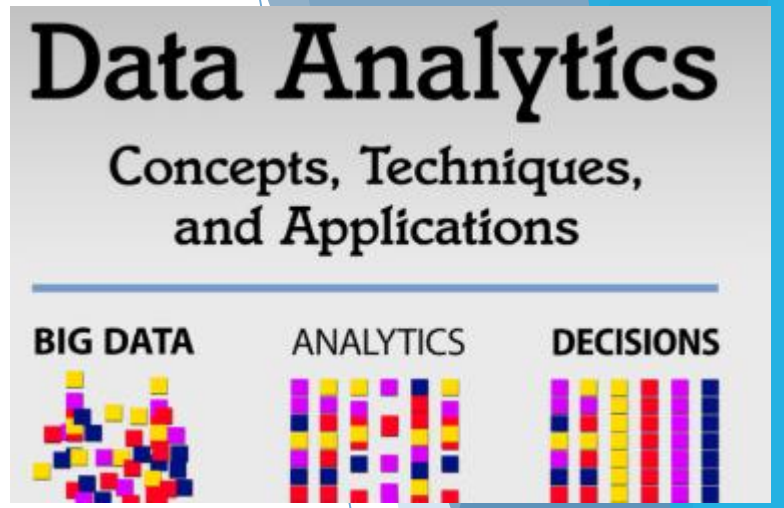

Traditional IT
Application
Data
Runtime
Middleware
OS
Virtualization
Servers
Storage
Networking
You manage
IaaS
Application
Data
Runtime
Middleware
OS
Virtualization
Servers
Storage
Networking
You manage
PaaS
Application
Data
Runtime
Middleware
OS
Virtualization
Servers
Storage
Networking
You manage
Serverless
Application
Data
Runtime
Middleware
OS
Virtualization
Servers
Storage
Networking
SaaS
Application
Data
Runtime
Middleware
OS
Virtualization
Servers
Storage
Networking

Data Analysis

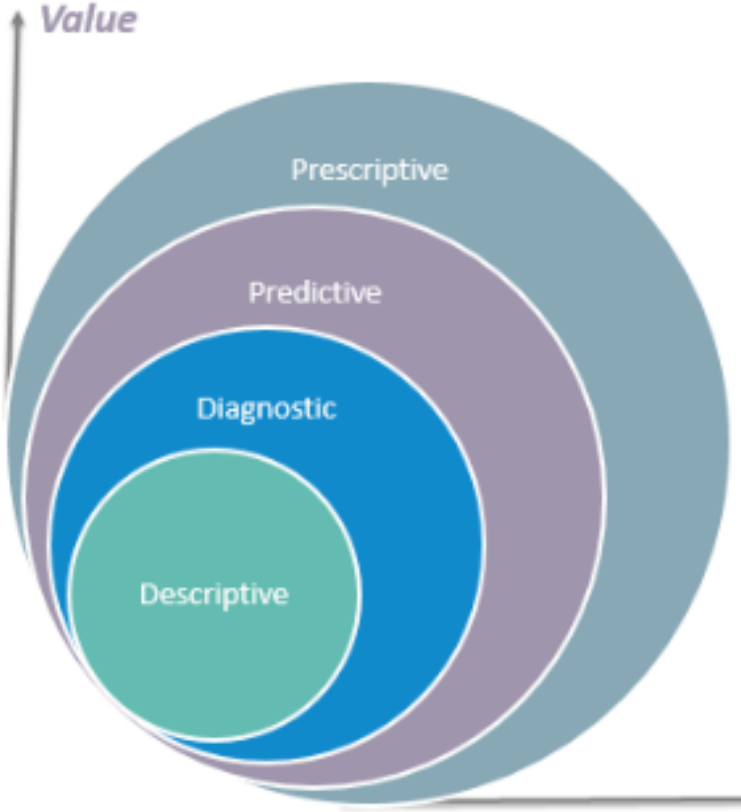
- Analysis is a detailed examination of something in order to understand its nature or determine its essential features.
- Data analysis is the process of compiling, processing, and analyzing data so that you can use it to make decisions.

Data Analytics

- Analytics is the systematic analysis of data
- Data analytics is the specific analytical process being applied



4 types of Data Analytics



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Complexity

Big Data

What is BIG DATA

- Big data involves the data produced by different devices and applications.
- A lot of data more than can easily be handled by a single database, computer or spreadsheet.
- Different kind of information in each record lacking inherent structure or predictable size, rate of arrival and transformation
- Ex:

  Black Box Data

  Social Media Data

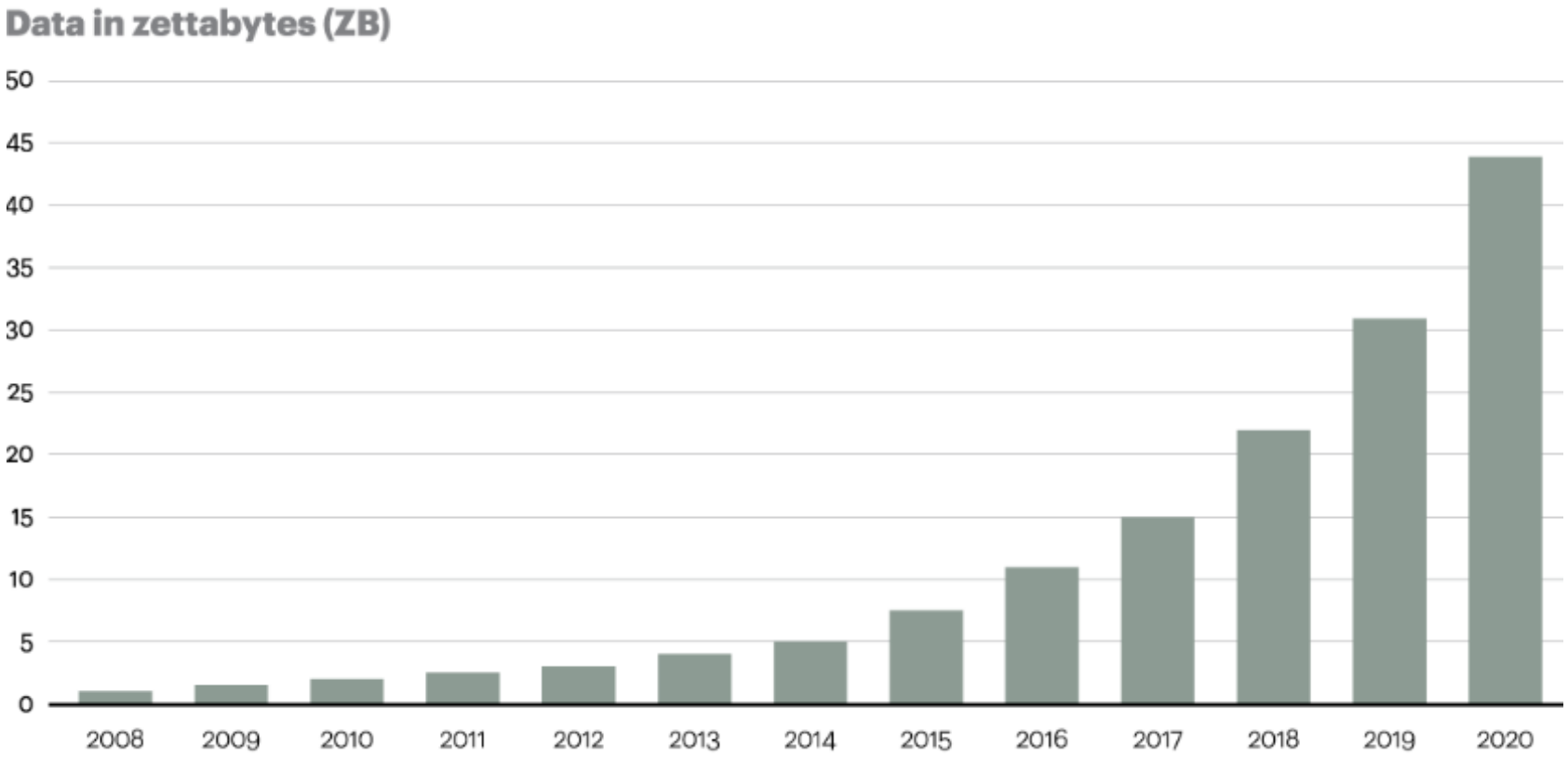
  Stock Exchange Data

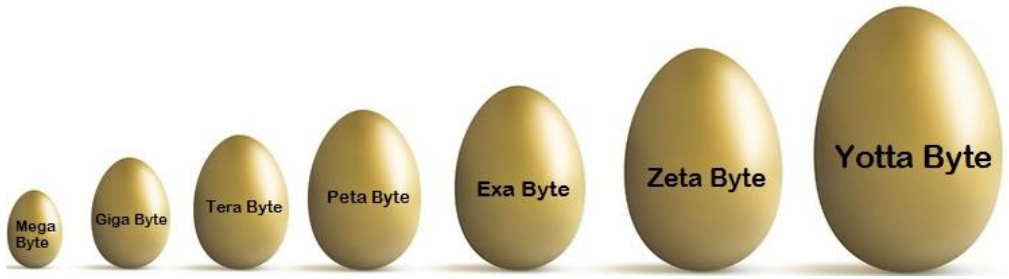
  Power Grid Data

  Transport Data

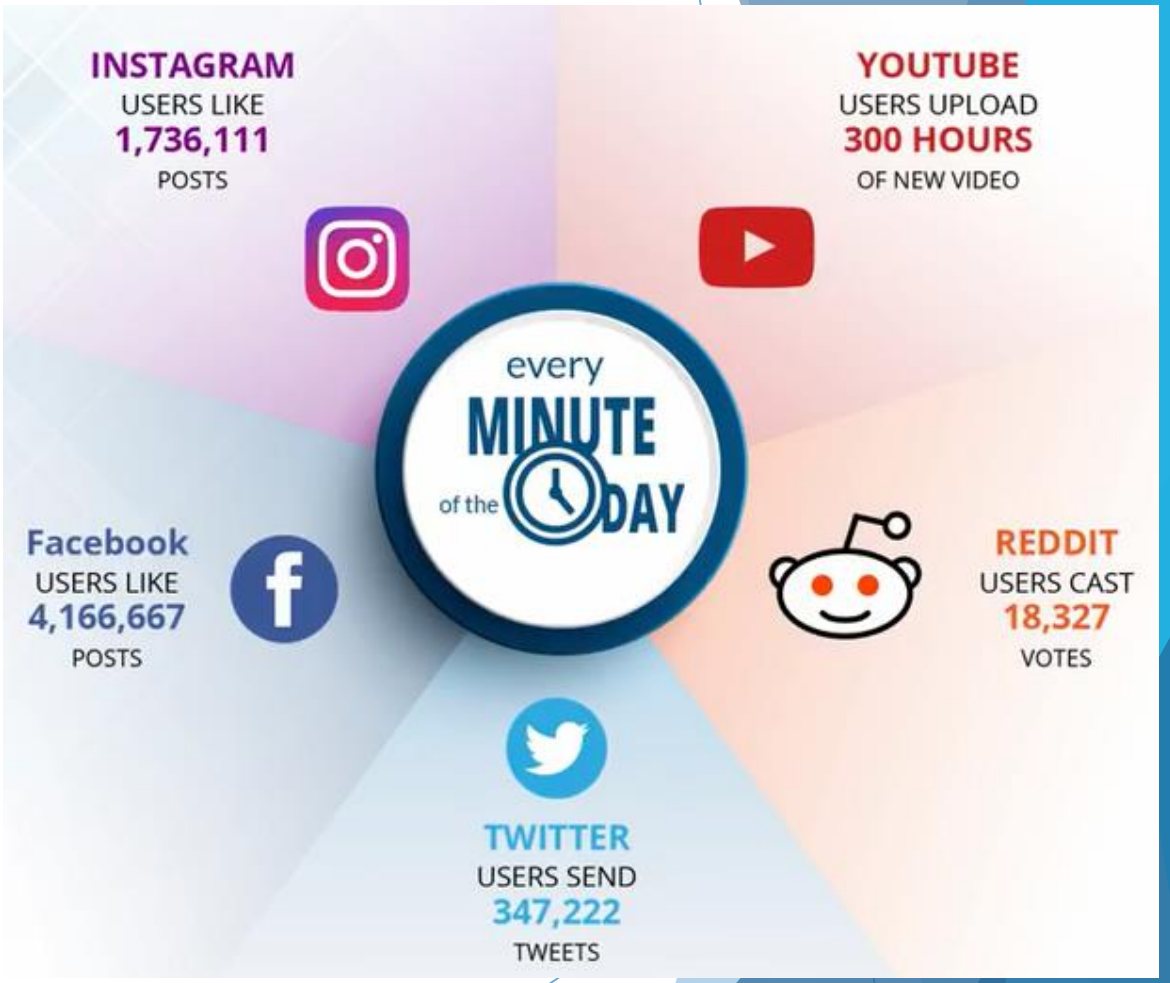
  Search Engine Data

Volume


Data in zettabytes (ZB)
50
45
40
35
30
25
20
15
10
5
0
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

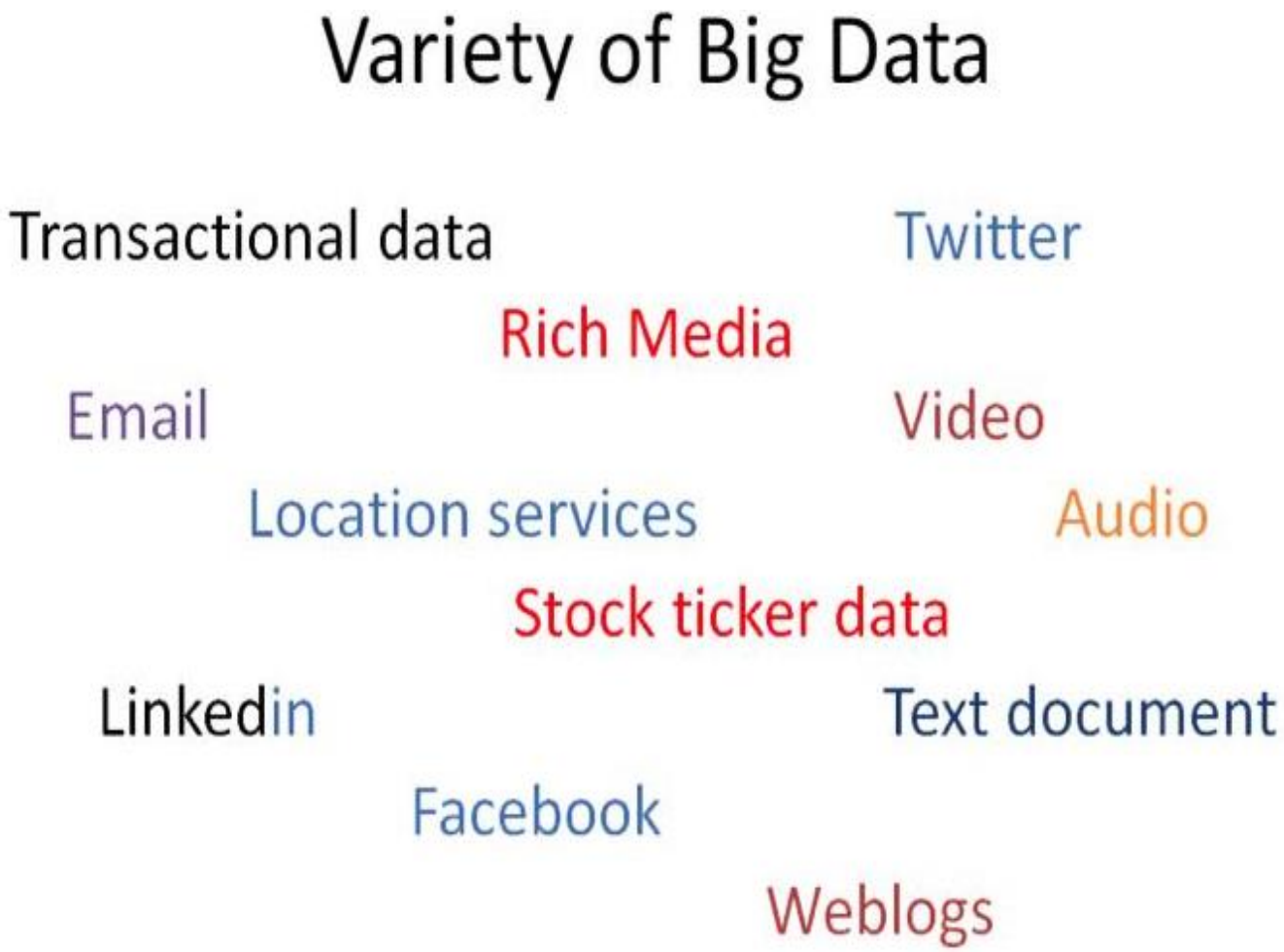

Mega Byte
Giga Byte
Tera Byte
Peta Byte
Exa Byte
Zeta Byte
Yotta Byte

Velocity



Airbus generates 10 TB every 30 minutes
About 640 TB is generated in one flight

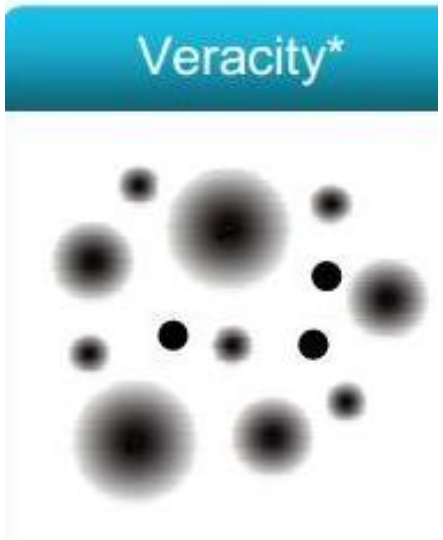
Variety


Variety of Big Data
Transactional data
Twitter
Rich Media
Email
Video
Location services
Audio
Stock ticker data
Linkedin
Text document
Facebook
Weblogs

Veracity

- Measurement error in the case of sensors?
- Should you trust all tweets about a given company?
- Lack of credential in case of social media

- Veracity provides confidence in the trustworthiness of the data



Hadoop

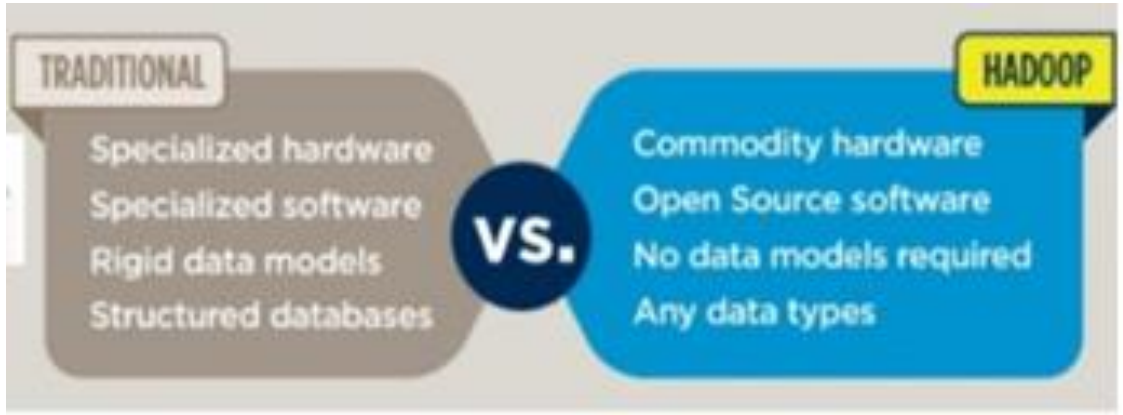
What is Hadoop

- Open source framework for distributed processing to handle Big Data by Apache
- Developed by Doug Cutting with support from Yahoo later
- Inspired by Map Reduce and Google File System where data is stored in commodity hardware with multiple replication to provide High Availability
- Designed to be scalable from a single system to support thousands of nodes

- **Yahoo has 4500 node Hadoop Cluster**
- **Facebook has 1100 node Hadoop Cluster**

The cute li'l yellow elephant is actually Doug's son's toy elephant; Hadoop is named after it!

Why Hadoop



- All the processing in Hadoop will be done in individual data nodes
- Scale Out is much more effective than Scale In
- Hadoop is open source low cost software

Hadoop is :

- Reliable
- Scalable
- Distributed computing

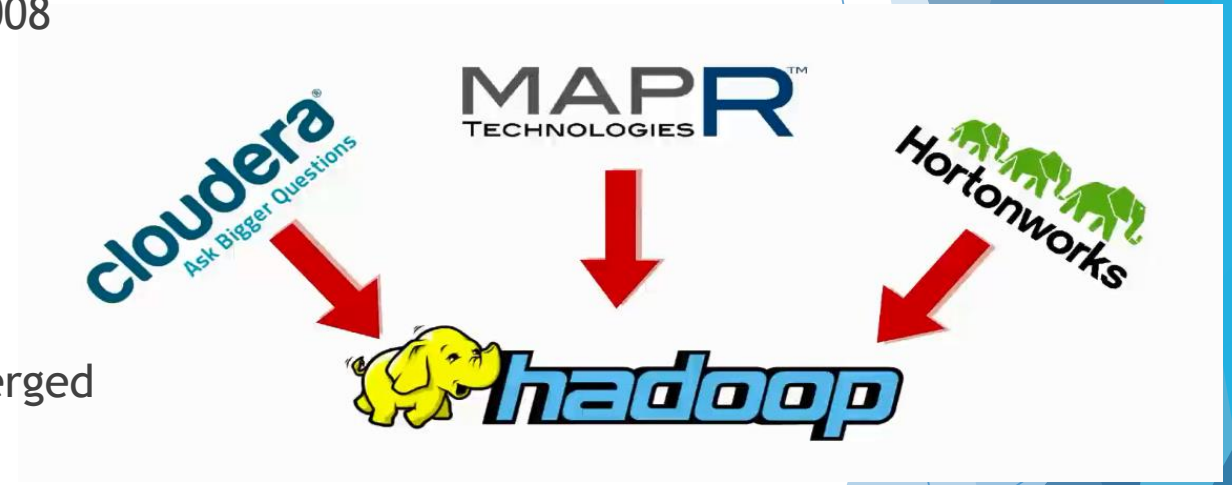
Hadoop Ecosystem

Hadoop – Distributed Networks

- They provide enterprise-ready Hadoop distributions

- Cloudera Inc. was founded by big data geniuses from Facebook, Google, Oracle and Yahoo in 2008 and is mostly used.
- Cloudera Manager is proprietary management software

  Versions: CDH 4.7.x, CDH 5.1.x

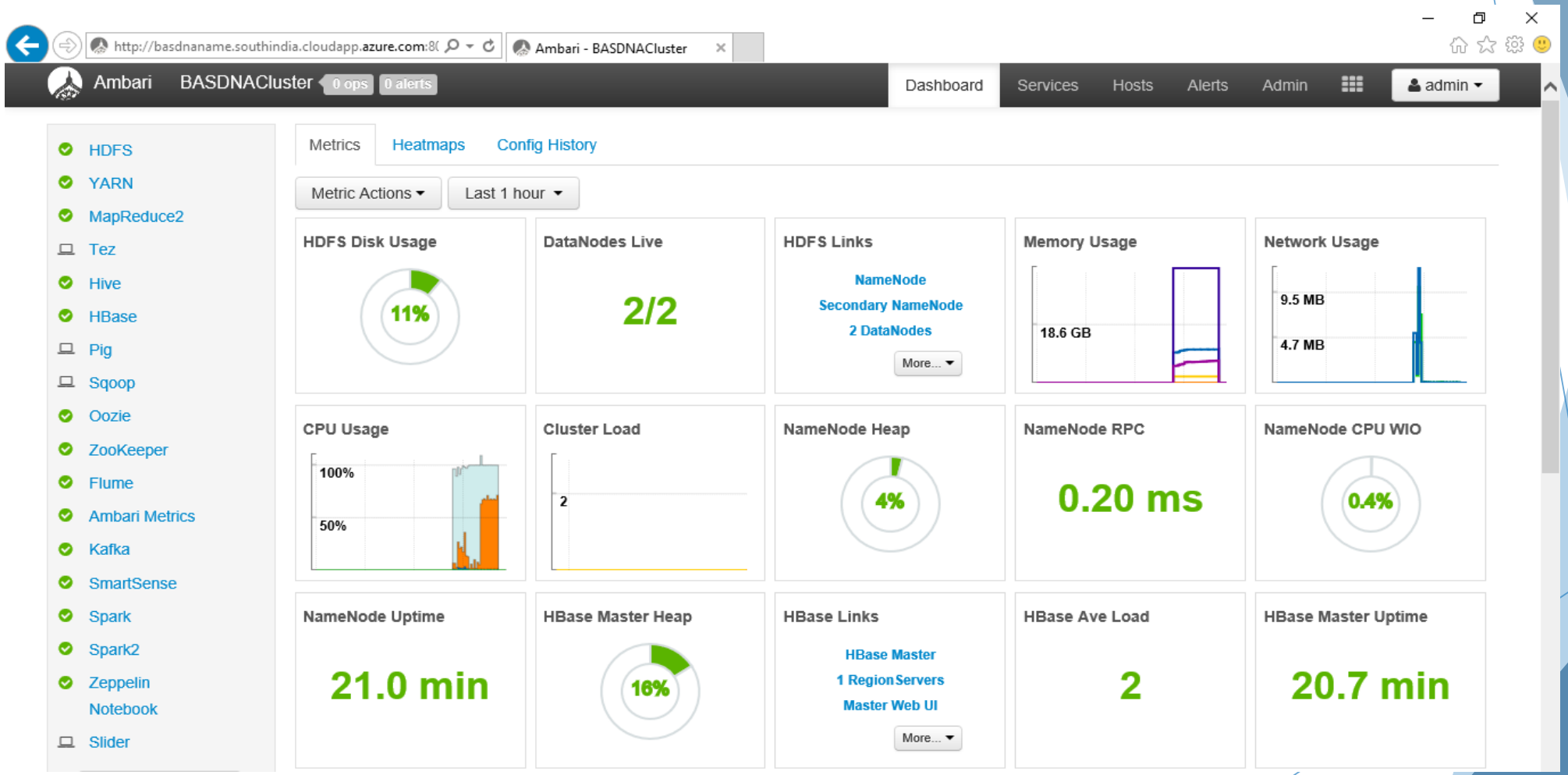


- Hortonworks, founded in 2011, has quickly emerged as one of the leading vendors of Hadoop
- Ambari is open source management software

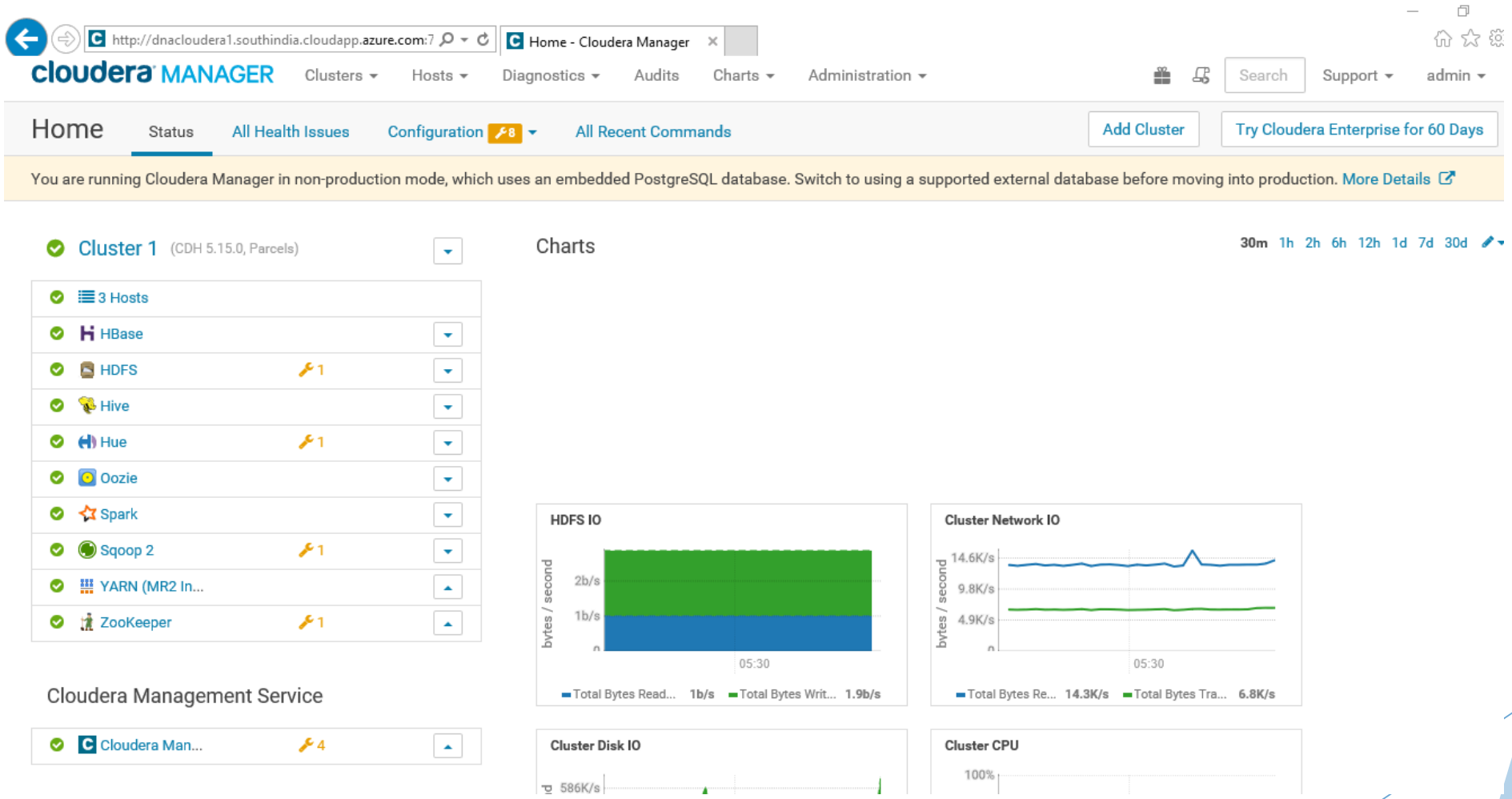
  Versions: HDP 2.1, HDP 2.2

- MapR develops and sells Apache Hadoop-derived software (Ex: MapRFS in the place of HDFS)

Hortonworks – Hadoop Components



Cloudera – Hadoop Components



Hadoop Cluster

Cluster is

- a set of connected computers that work together
- For storing and analyzing huge amounts of structured and unstructured data

- Each computer is called as Node
- Additional nodes can be added to a cluster

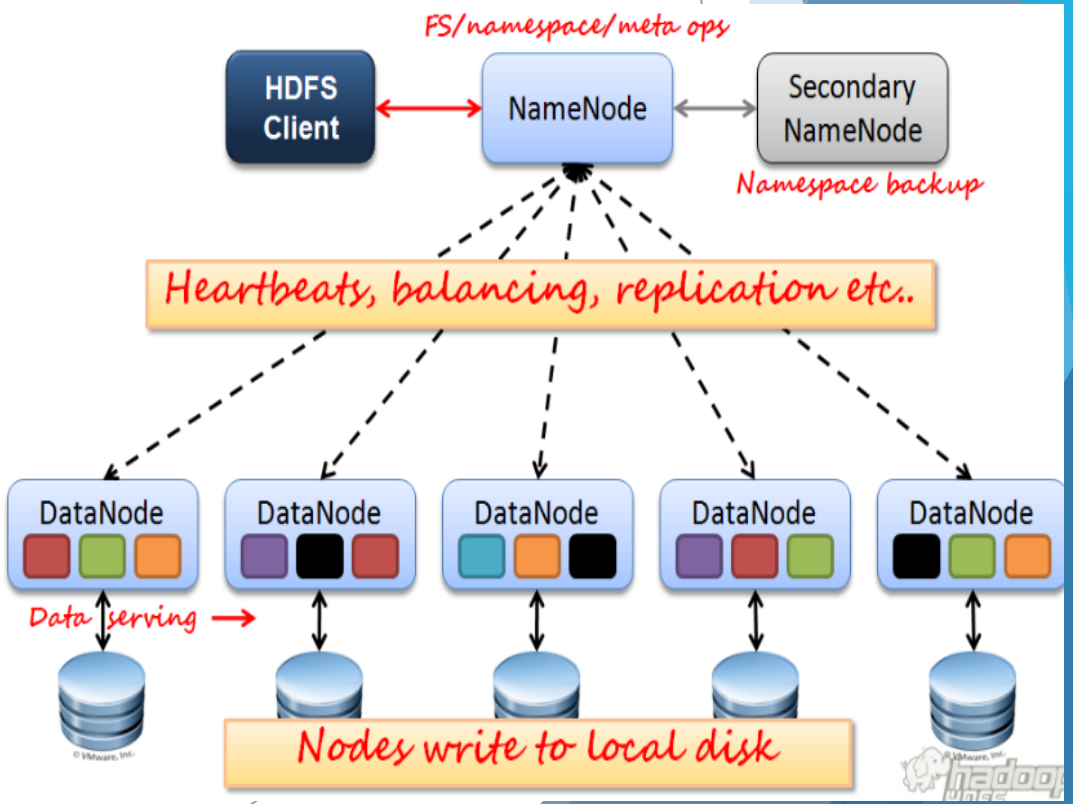
- The two components of HDFS cluster are:

Name Node

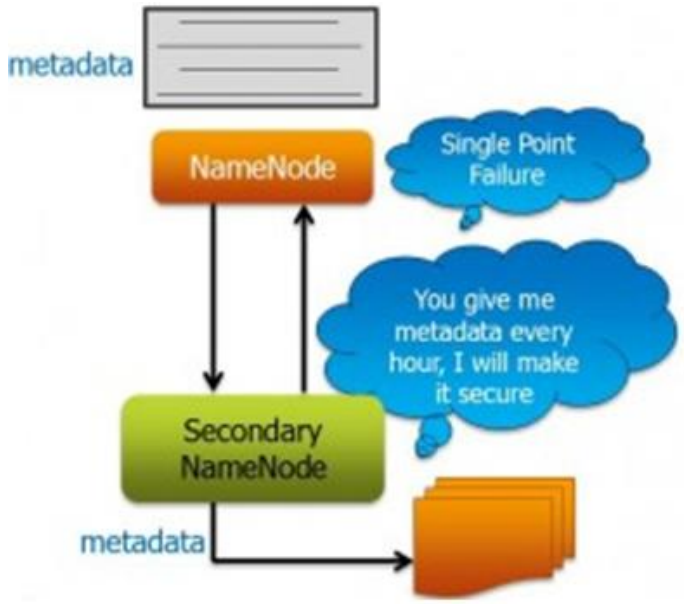
- The NameNode determines the mapping of blocks to DataNodes.

Data Node

- Sending heartbeats to the NameNode
- Sending a Blockreport to the NameNode



Secondary NameNode



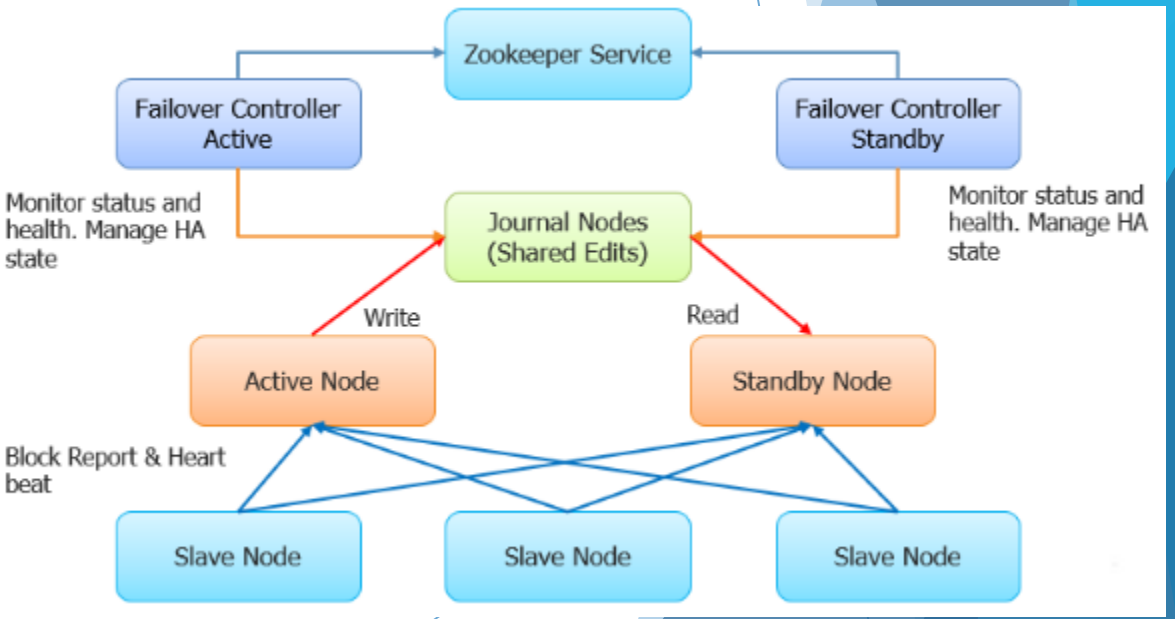
Secondary NameNode:

- Connects to NameNode regularly
- Housekeeping, backup of NameNode metadata
- Saved metadata can build a failed NameNode

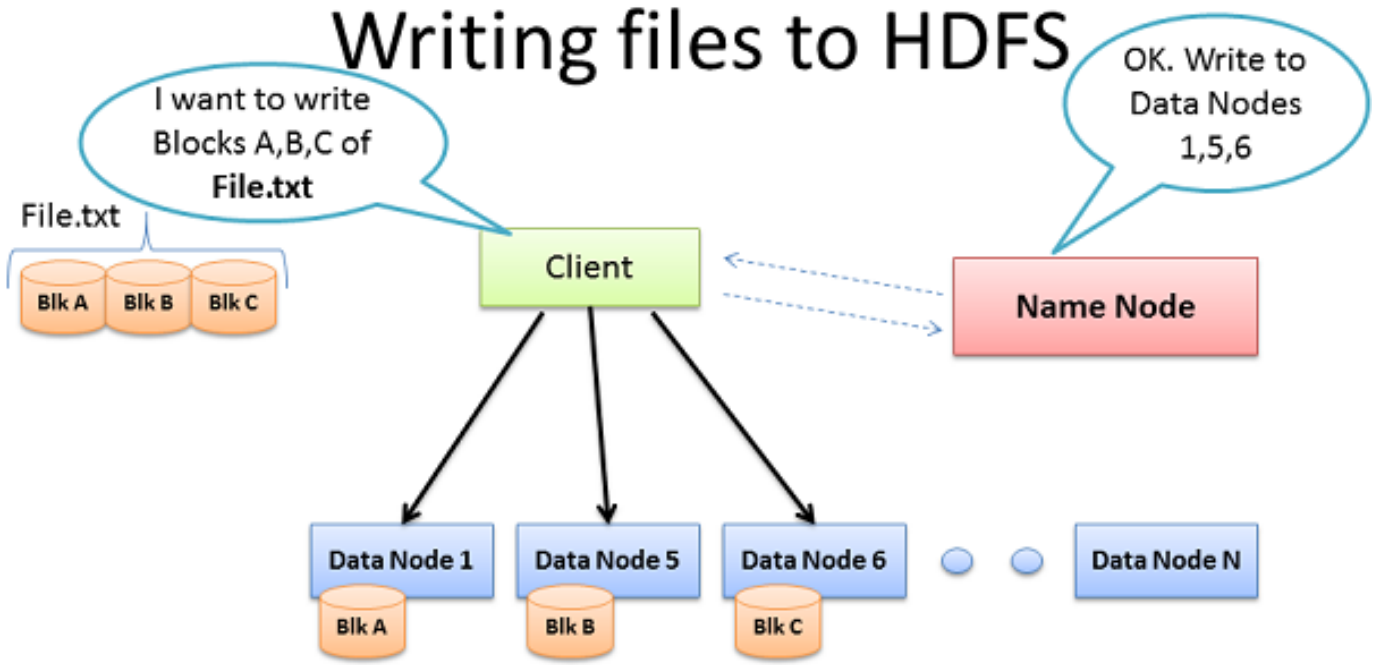
```
[hduser@Machine1 hadoop]$ ls -all /keermetadata/name/current
total 13812
drwxr-xr-x. 2 hduser hadoop   12288 Sep 22 18:28 .
drwxr-xr-x. 3 hduser hadoop    4096 Sep 20 15:17 ..
-rw-r--r--. 1 hduser hadoop 1048576 Jul 17 19:44 edits_0000000000000000001-0000000000000000001
-rw-r--r--. 1 hduser hadoop      42 Jul 18 20:12 edits_0000000000000000002-0000000000000000003
-rw-r--r--. 1 hduser hadoop 1048576 Jul 18 20:12 edits_0000000000000000004-0000000000000000004
-rw-r--r--. 1 hduser hadoop 1048576 Jul 19 23:46 edits_0000000000000000005-0000000000000000011
-rw-r--r--. 1 hduser hadoop      42 Jul 20 15:47 edits_0000000000000000012-0000000000000000013
-rw-r--r--. 1 hduser hadoop      42 Jul 20 16:47 edits_0000000000000000014-0000000000000000015
-rw-r--r--. 1 hduser hadoop 1048576 Jul 20 16:47 edits_0000000000000000016-0000000000000000016
-rw-r--r--. 1 hduser hadoop      42 Aug 18 00:37 edits_0000000000000000017-0000000000000000018
-rw-r--r--. 1 hduser hadoop      42 Aug 18 01:37 edits_0000000000000000019-0000000000000000020
-rw-r--r--. 1 hduser hadoop      42 Aug 18 02:37 edits_0000000000000000021-0000000000000000022
-rw-r--r--. 1 hduser hadoop      42 Aug 18 03:37 edits_0000000000000000023-0000000000000000024
-rw-r--r--. 1 hduser hadoop      42 Aug 18 04:37 edits_0000000000000000025-0000000000000000026
-rw-r--r--. 1 hduser hadoop     570 Sep 22 18:28 edits_0000000000000000295-0000000000000000303
-rw-r--r--. 1 hduser hadoop     570 Sep 22 18:28 edits_0000000000000000295-0000000000000000303
-rw-r--r--. 1 hduser hadoop 1048576 Sep 22 18:28 edits_inprogress_0000000000000000304
-rw-r--r--. 1 hduser hadoop    1143 Sep 22 18:18 fsimage_0000000000000000294
-rw-r--r--. 1 hduser hadoop      62 Sep 22 18:18 fsimage_0000000000000000294.md5
-rw-r--r--. 1 hduser hadoop    1143 Sep 22 18:28 fsimage_0000000000000000303
-rw-r--r--. 1 hduser hadoop      62 Sep 22 18:28 fsimage_0000000000000000303.md5
-rw-r--r--. 1 hduser hadoop       4 Sep 22 18:28 seen_txid
-rw-r--r--. 1 hduser hadoop     202 Sep 22 14:55 VERSION
```

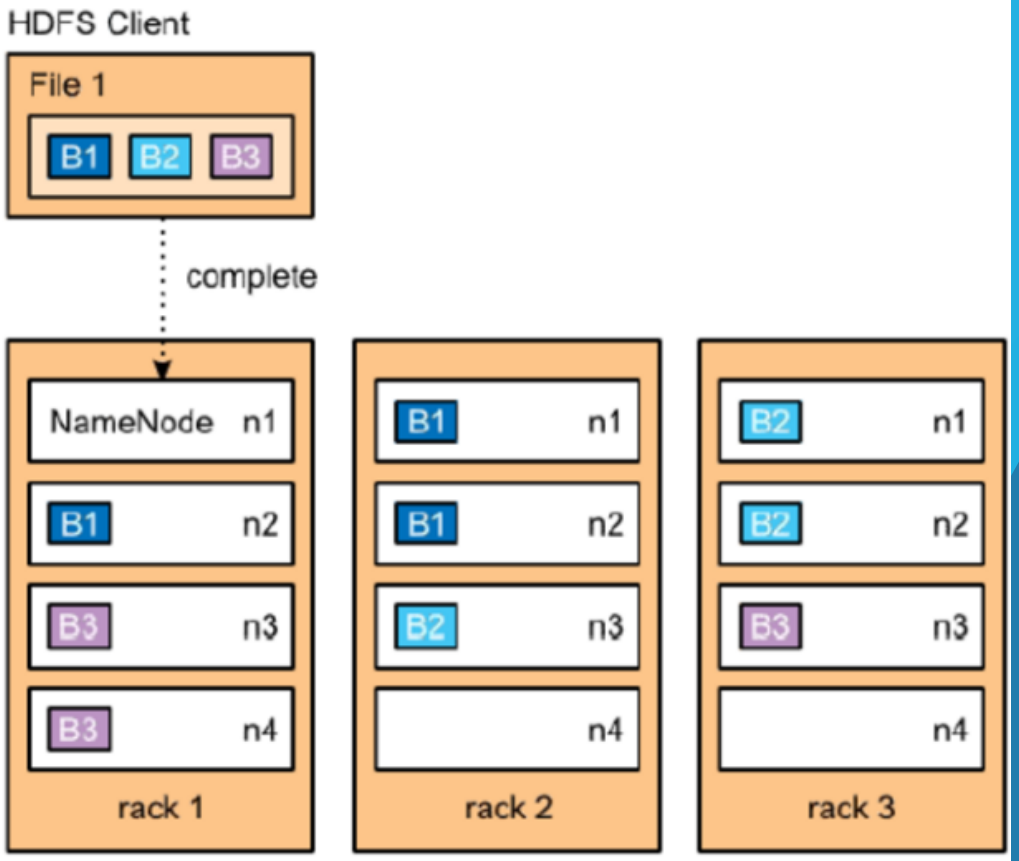
High Availability

- Two separate machines are configured as NameNodes
- One in Active state and other in a Standby state
- Active NameNode is responsible for all client operations in the cluster
- Both nodes communicate with a group of separate daemons called ‘JournalNodes’ (JNs)
- Active node logs record of the changes made in the JournalNodes.
- Standby node read the amended information from the JNs, and is regularly monitoring them for changes.
- In case of a failover, the Standby read all the changes from the JounalNodes before changing its state to ‘Active state’.
- DataNodes send block location information and heartbeats to both.
- During a failover, the NameNode which is to become active will take over the responsibility of writing to the JournalNodes.
- Zookeeper is used for automatic failover



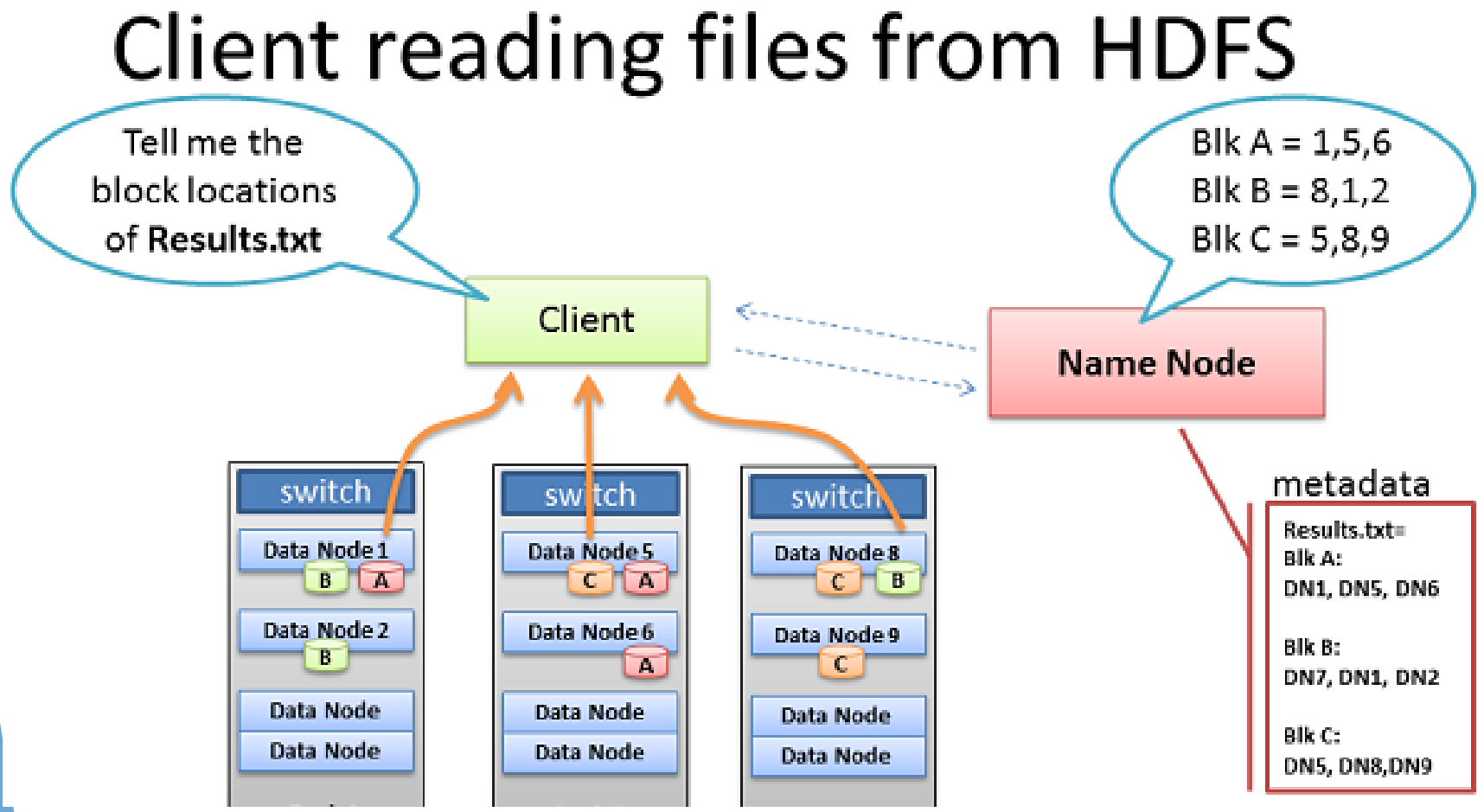
Writing in HDFS





- The default block size is 128MB
- The default replication factor is three

Reading from HDFS



HDFS Commands

HDFS – Command Line Interface (hadoop fs command)

Get a directory listing in users home directory in HDFS – (if home directory exists)

- hadoop fs -ls

Get a directory listing of the HDFS root directory

- hadoop fs –ls /

Creating a directory in HDFS

- hadoop fs –mkdir <dirname>

Copying a local file into HDFS

- hadoop fs –copyFromLocal <local fileName> <targetDirectory/targetFileName>
- hadoop fs –put <fileName> <targetDirectory/targetFileName>

Display the contents of a file

- hadoop fs –cat <filename>

Copying a HDFS file into local file system

- hadoop fs –copyToLocal <sourceDirectory/sourceFileName> <local fileName>
- hadoop fs –get <sourceDirectory/sourceFileName> <local fileName>

Removing a directory and its contents from HDFS

- hadoop fs –rmr <targetDirectory>

About MapReduce

- To process large amounts of data in parallel across a distributed environment.
- A MapReduce program consists of two main phases:

Map phase :

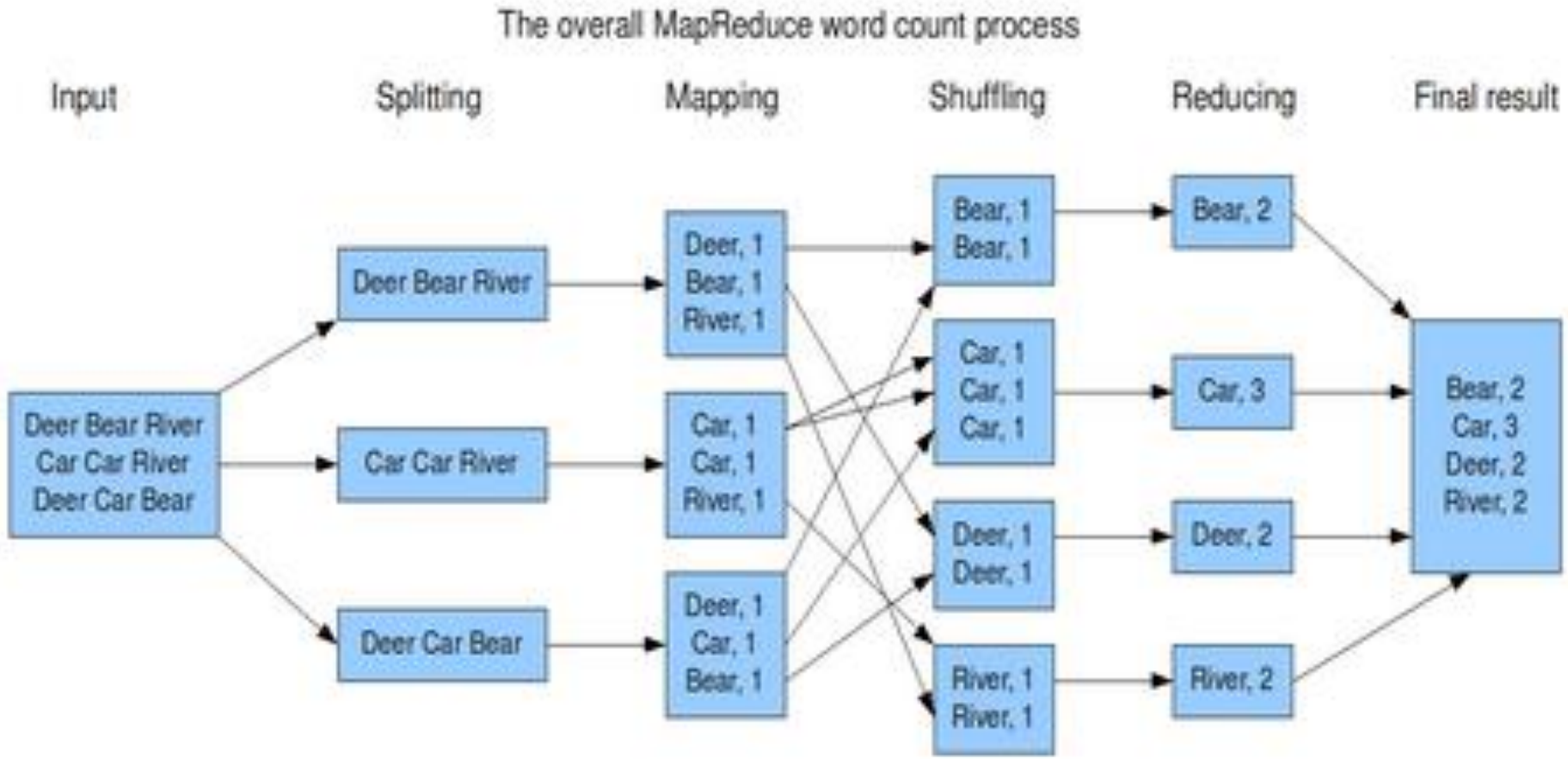
- Data is input into the mapper, where it is transformed and prepared for the reducer
- The mapper inputs key/value pairs from HDFS files and outputs intermediate key/value pairs

Reduce phase:

- Retrieves the data from the mapper and performs the desired computations or analyses

- The number of mappers is determined by the input format
- The number of reducers is determined by the MapReduce job configuration

Map Reduce - Example



YARN

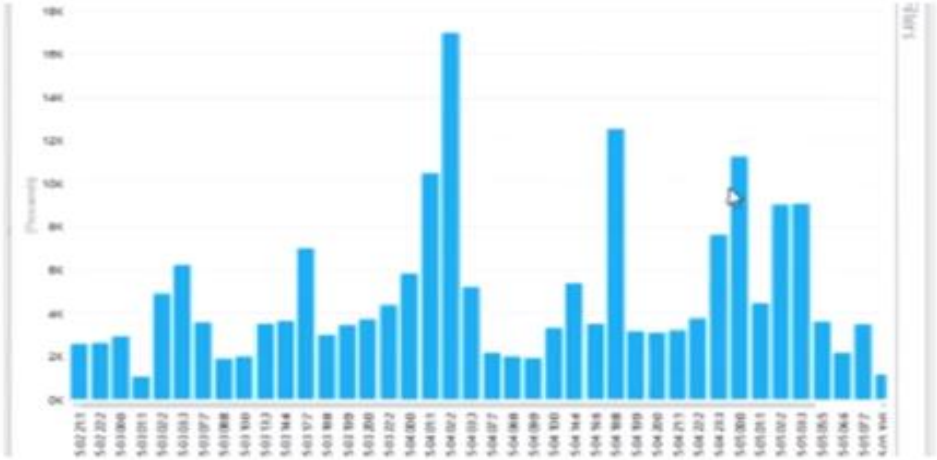
- YARN is Yet Another Resource Negotiator
- YARN provides better resource management in Hadoop, resulting in improved cluster efficiency

Sentiment Use Case

- Tracking the volume of tweets around the movie's launch

Sentiment Use Case

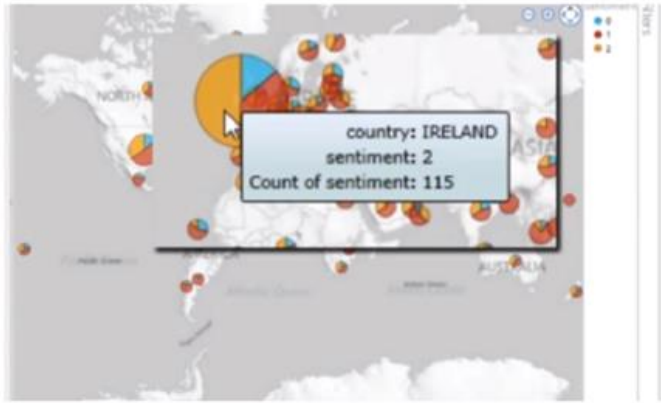
- Analyze customer sentiment on the days leading up to and following the release of the movie *Iron Man 3*.
- Questions to answer:
  - How did the public feel about the debut?
  - How might the sentiment data have been used to better promote the launch of the movie?



Notice a large spike in tweets around the Thursday midnight opening, and spikes around the Friday evening, Saturday afternoon and Saturday evening showings.

View Spikes in Tweet Volume

The sentiment of the tweets was graphed by country:



Viewing the tweets on a map shows the sentiment of the movie by country. For example, Ireland had 50% positive tweets, while 67% of tweets from Mexico were neutral.

View Sentiment by Country

Clickstream Use Case

- Amazon's use of real-time, item-based, collaborative filtering (IBCF) to fuel its 'Frequently bought together' and 'Customers who bought this item also bought' features
- Amazon generates about 20% more revenue via this method.

- LinkedIn and Facebook suggesting 'People you may know' or 'Companies you may want to follow'.

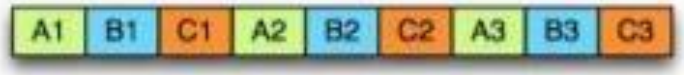
Columnar Format

- Better Compression as data is more homogeneous
- We can efficiently scan only a subset of columns while reading the data
- Works well for queries that only access a small subset of columns
- Format Ex: Optimized Row Columnar (ORC) and Parquet

To illustrate what columnar storage is all about, here is an example with three columns.

A	B	C
A1	B1	C1
A2	B2	C2
A3	B3	C3

In a row-oriented storage, the data is laid out one row at a time as follows:



Whereas in a column-oriented storage, it is laid out one column at a time:

File Format

- Standard file format consists of Text files, CSV, XML or binary file types (such as images).
- SequenceFiles store data as binary key-value pairs
- SerDe is short for serializer/deserializer and refers to how records read in from a table (deserialized) and written back out to HDFS (serialized).

- Avro - write-heavy
  - Row-based storage format
  - Contains its own schema and schema evolution
  - Amenable to "full-table scans"
- Parquet - read-heavy
  - Columnar storage formats, sometimes used for final storage
  - Smaller disk-reads
  - Great for feature selection
- ORC - read-heavy
  - Defaults to Zlib compression
  - Mixed row-column, splittable

Compression

- Compression is important consideration for storing data in Hadoop for reducing storage requirements and improving data processing performance
- Splittability is a major consideration in choosing a compression format as well as file format
- Any compression format can be made splittable when used with container file formats like Avro, SequenceFiles

Format	Strengths	Weakness
Snappy	• Developed at Google for high compression speeds	• Doesn't offer the best compression sizes • Relatively slow in decompression • Non-Splittable
LZO	• Rapid compression • Balanced compression and decompression times	• Doesn't offer the best compression sizes • Non-Splittable - can be made splittable by adding additional indexing step • LZO's license requires separate install to be distributed within hadoop
Gzip	• Good compression performance • Reasonable speed • Smaller blocks with Gzip can lead to better performance	• Relatively slower in write speed than snappy and LZ4 • Non-Splittable
bzip2	• Excellent compression performance • Splittable	• bzip2 is about 10 times slower than Gzip
LZ4	• Quick Compression • Best results in decompression	• Non-Splittable

Azure

Azure Synapse Analytics

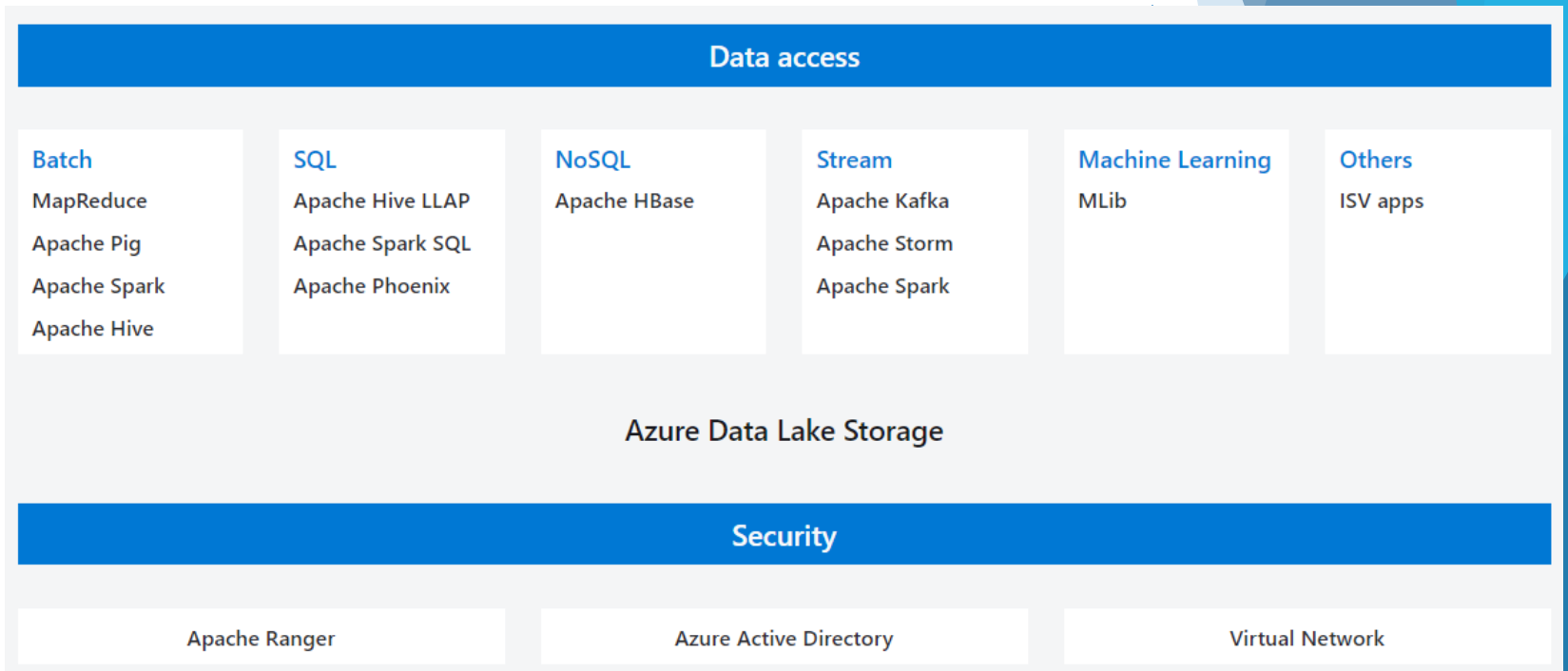
- Azure Synapse is an analytics service that brings together enterprise data warehousing and Big Data analytics.
- Azure SQL Data Warehouse uses distributed data and a massively parallel processing (MPP) design.

- Synapse SQL pool represents a collection of analytic resources that are being provisioned when using Synapse SQL.
- The size of SQL pool is determined by Data Warehousing Units (DWU).
- PolyBase uses standard T-SQL queries to bring the data into Synapse SQL pool tables.
- Synapse SQL pool stores data in relational tables with columnar storage
- This format significantly reduces the data storage costs, and improves query performance.

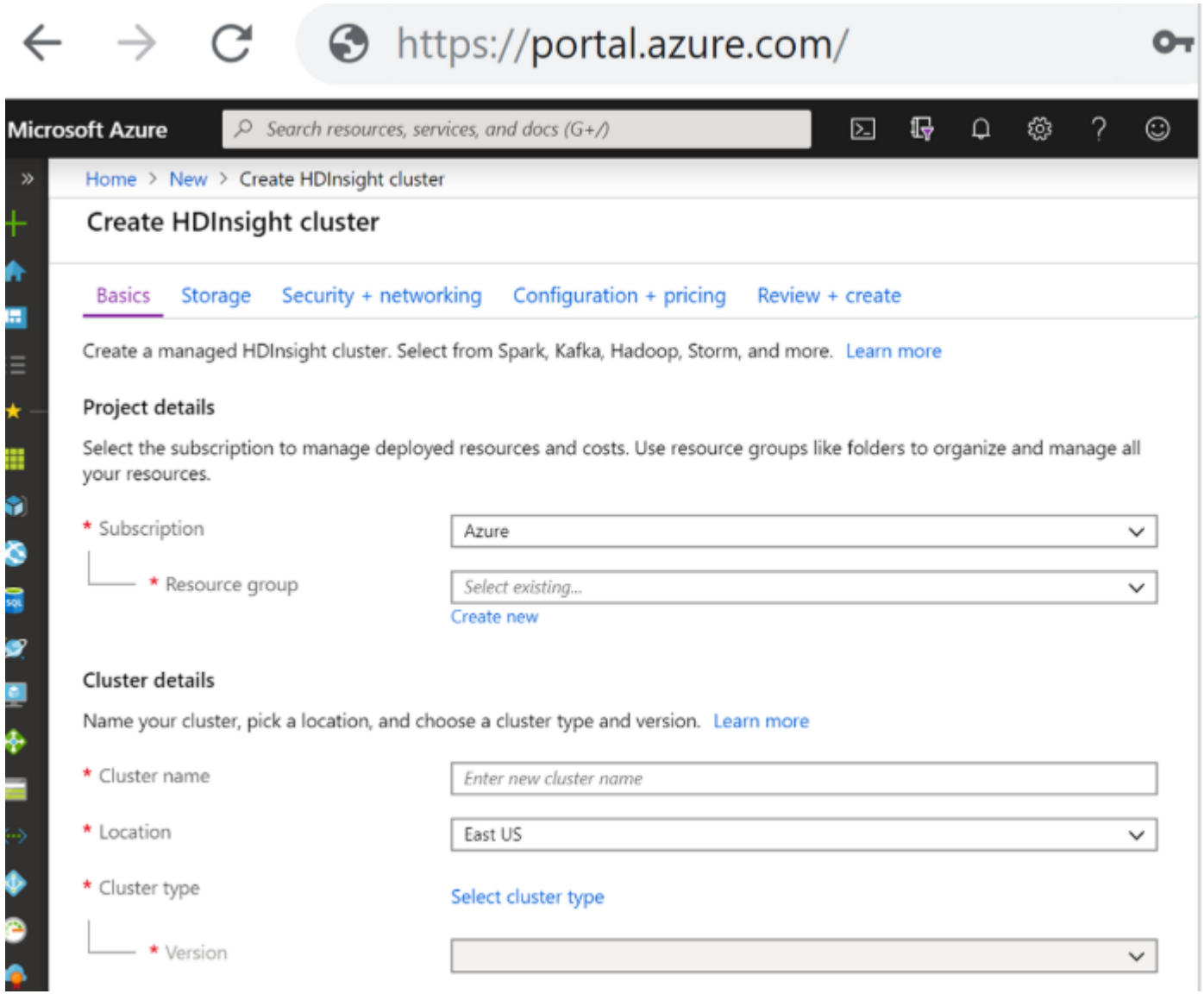
Azure HDInsight

- HDInsight offers the following cluster types

- HDInsight offers the following cluster types
- Azure HDInsight is a cloud distribution of Hadoop components.
- Azure HDInsight can be used for a variety of scenarios in big data processing.



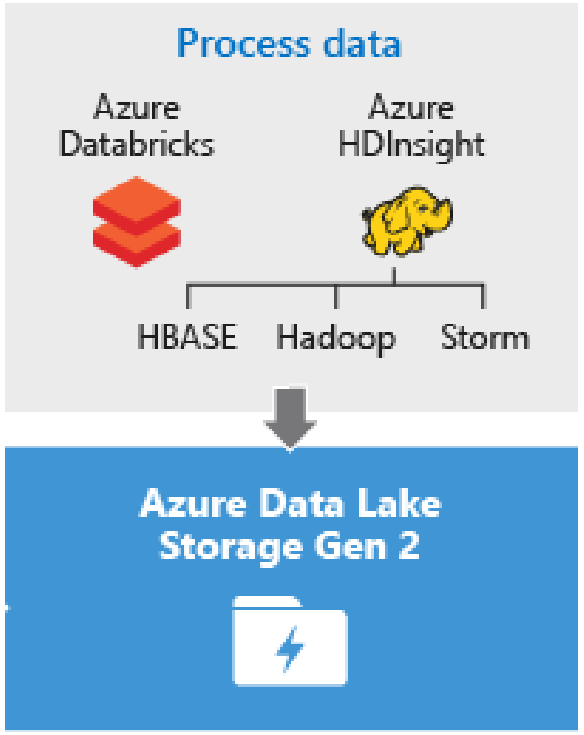
HDInsight


https://portal.azure.com/
Microsoft Azure
Search resources, services, and docs (G+/)
Home > New > Create HDInsight cluster
Create HDInsight cluster
Basics
Storage
Security + networking
Configuration + pricing
Review + create
Create a managed HDInsight cluster. Select from Spark, Kafka, Hadoop, Storm, and more. Learn more
Project details
Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.
* Subscription
Azure
* Resource group
Select existing...
Create new
Cluster details
Name your cluster, pick a location, and choose a cluster type and version. Learn more
* Cluster name
Enter new cluster name
* Location
East US
* Cluster type
Select cluster type
* Version

Azure Data Lake

- Azure Data Lake includes all the capabilities required to make it easy for developers, data scientists, and analysts to store data of any size, shape, and speed, and do all types of processing and analytics across platforms and languages.
- Azure Data Lake Storage Gen2 is a highly scalable and cost-effective data lake solution for big data analytics

- The Data Lake store provides a single repository where organizations upload data of just about infinite volume.
- Azure Data Lake store introduced a new file system called AzureDataLakeFilesystem (adl://)



Thank You

Keerthiga Barathan