BigData – Advance

Abdul H Khan

ver. 1.0

Class & sessions agenda

Session 1 - Overview and Use Cases

- Quick overview of trends and use cases
- Tools and Eco-system projects we will cover
- Perform quick health check of Hadoop cluster
- Setup Nifi streaming engine
- Create - Setup Twitter application API

Session 2 - Ingestion and Data manipulation

- Setup Nifi ingestion stream info HDFS
- Create table on RAW twitter dataset
- convert Raw data into structure dataset
- Ingest data into HDFS via raw text file
- MapReduce sample program

Session 3 - Overview and Use Cases

- Install R and R studio server
- Install Rhadoop libraries
    - Data extraction using Hadoop Streaming
- Experiment with Twitter Data from HDFS
    - Find top user
    - Find top keywords
    - Find Positive and negative

Session 4 - Overview and Use Cases

- Why Spark?
- Install Spark on Yarn
- What is Spark-Submit, Spark-Shell and pyspark
- Spark toolkit in field - Jupyter notebook, Levy server, H20, SparkR
- Read data from existing file in HDFS and display 10 values
- Query existing Hive table via SparkSQL

Session 5 – Visual and Data analytics

- Install Solr service on HDFS cluster
    - Create Solr collection
    - Integrate Solr with Lucidworks Banana visual
    - Data visualization with static NYC taxi or AirBnB data

Typical use cases of BigData

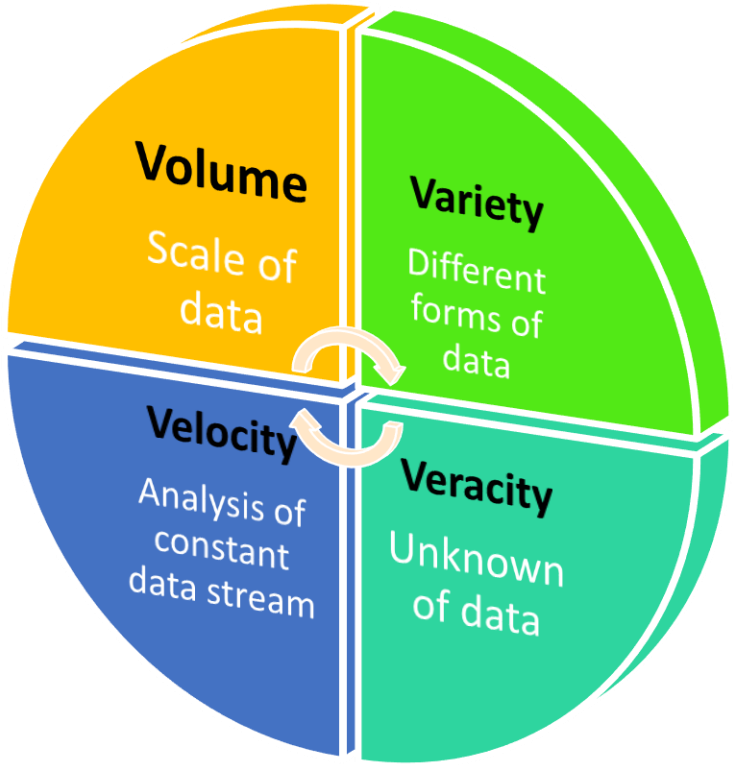
- What is Big Data?
  - Why use Big Data?
  - How can Big Data help your enterprise? Should you use Big Data?
  - Do you have Big Data problem?
- Industry typical use case for Big Data
  - Data Lake
  - Data offloading – Hybrid environment
  - Data analytics – Data discovery
  - Machine learning – Deep learning

What is Big Data?

- Extremely large datasets (Terabytes > Petabytes > Exabytes > Zettabytes)
- Exceeds the processing capacity of conventional database systems
- As far back as 2001, industry analyst Doug Laney (currently with Gartner) articulated the now mainstream definition of big data as the three Vs of big data: volume, velocity and variety[1].
  - **Volume.** Many factors contribute to the increase in data volume. Transaction-based data stored through the years. Unstructured data streaming in from social media. Increasing amounts of sensor and machine-to-machine data being collected. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.
  - **Velocity.** Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.
  - **Variety.** Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data is something many organizations still grapple with.

4 V’s

- Click Streaming data
- Sensors
- Log
- Events
- Speech
- Social media

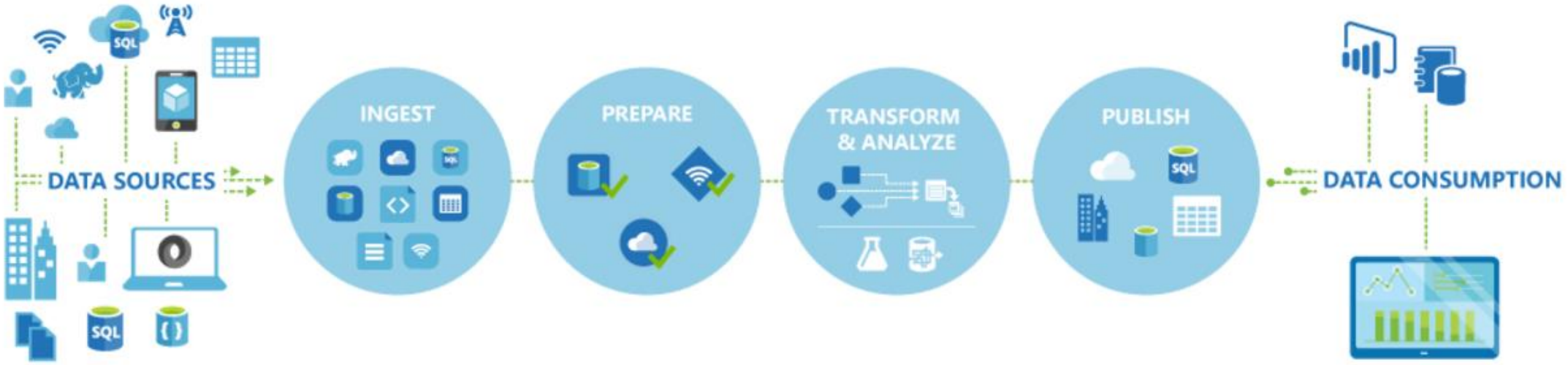


- Unstructured
- Semi-structured
- Structured

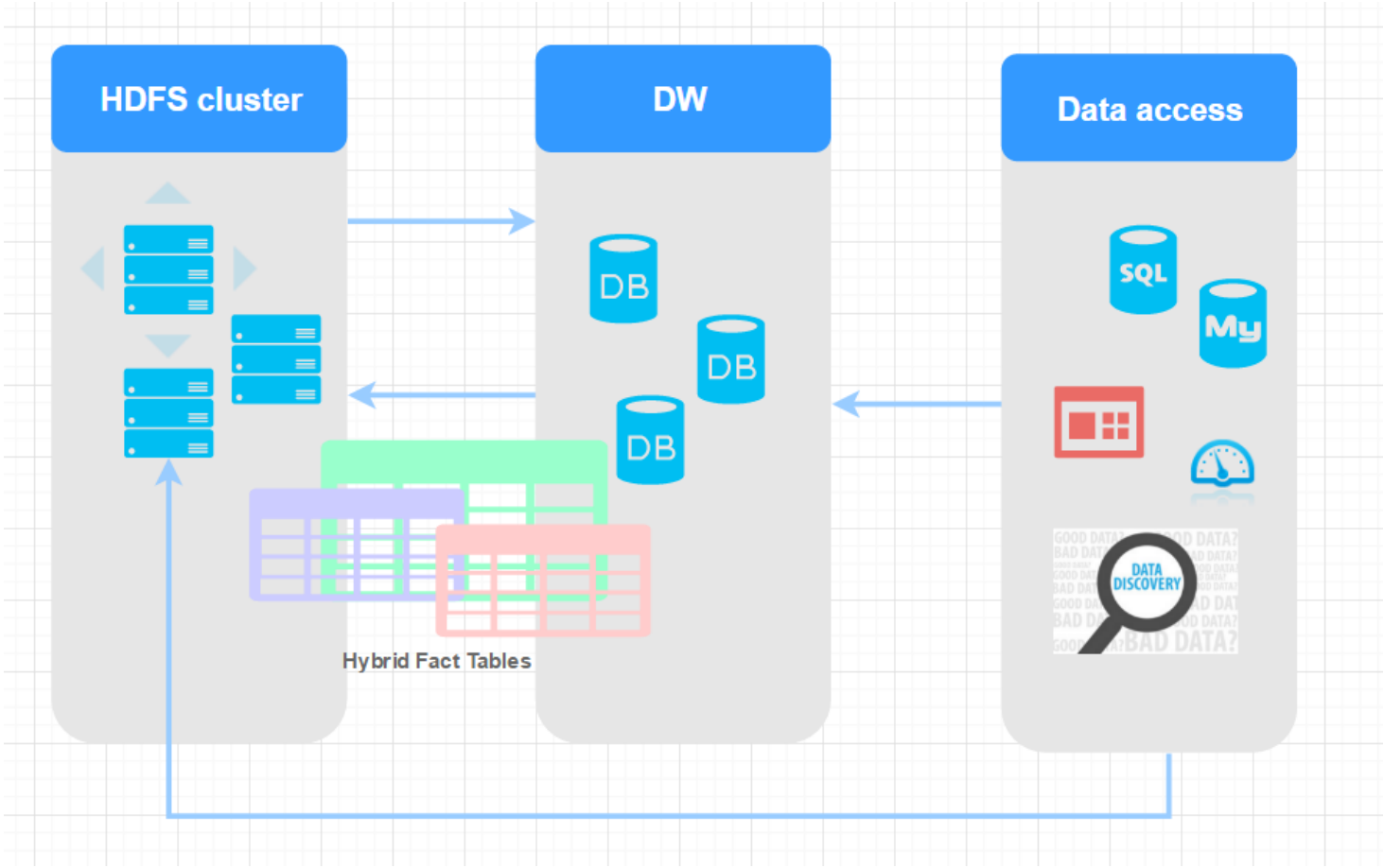
- Speed of data generation
- Rate of analysis

- Untrusted
- Raw
- Uncleansed

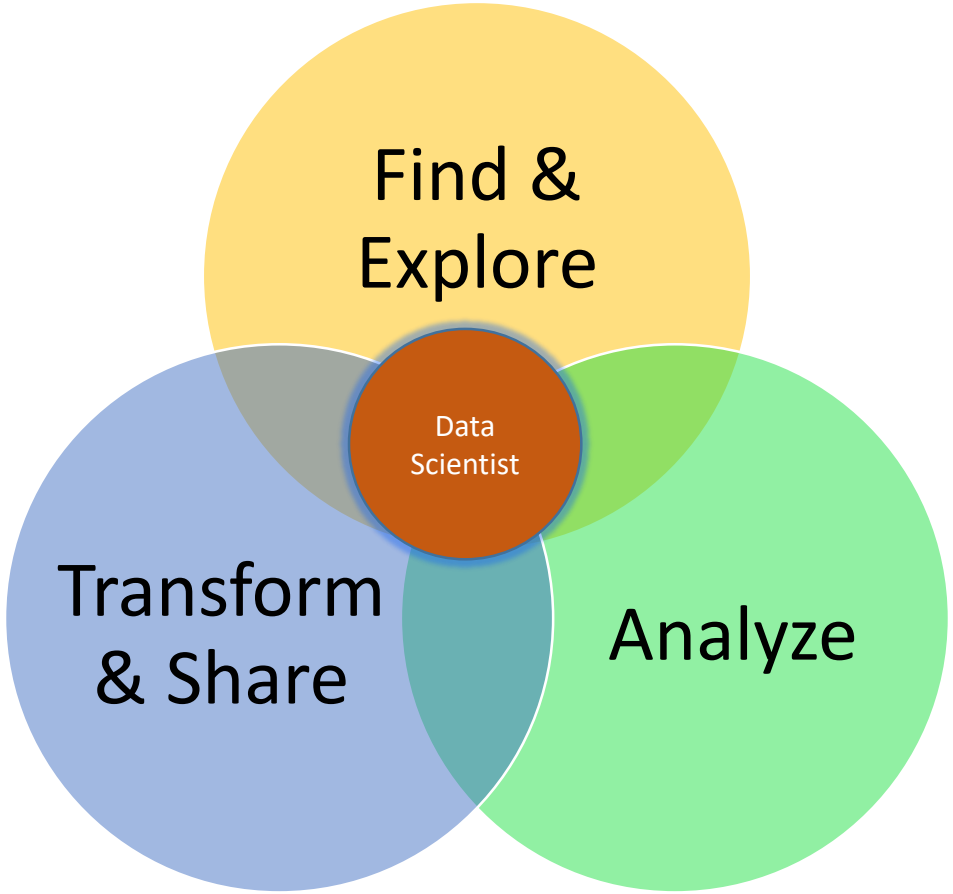
Data Lake or Data Factory



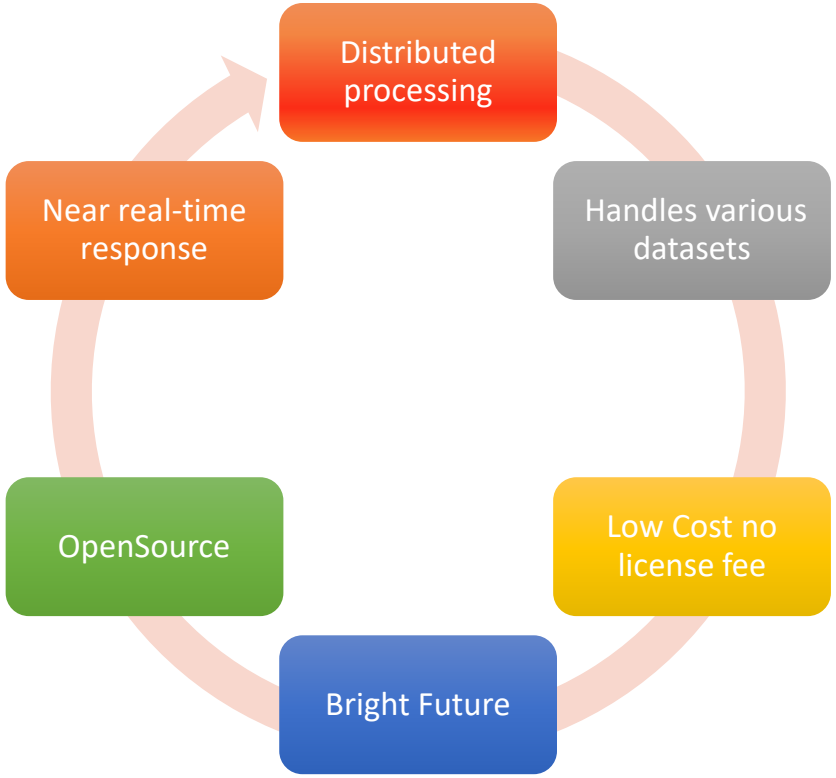
Data offloading – Hybrid environment

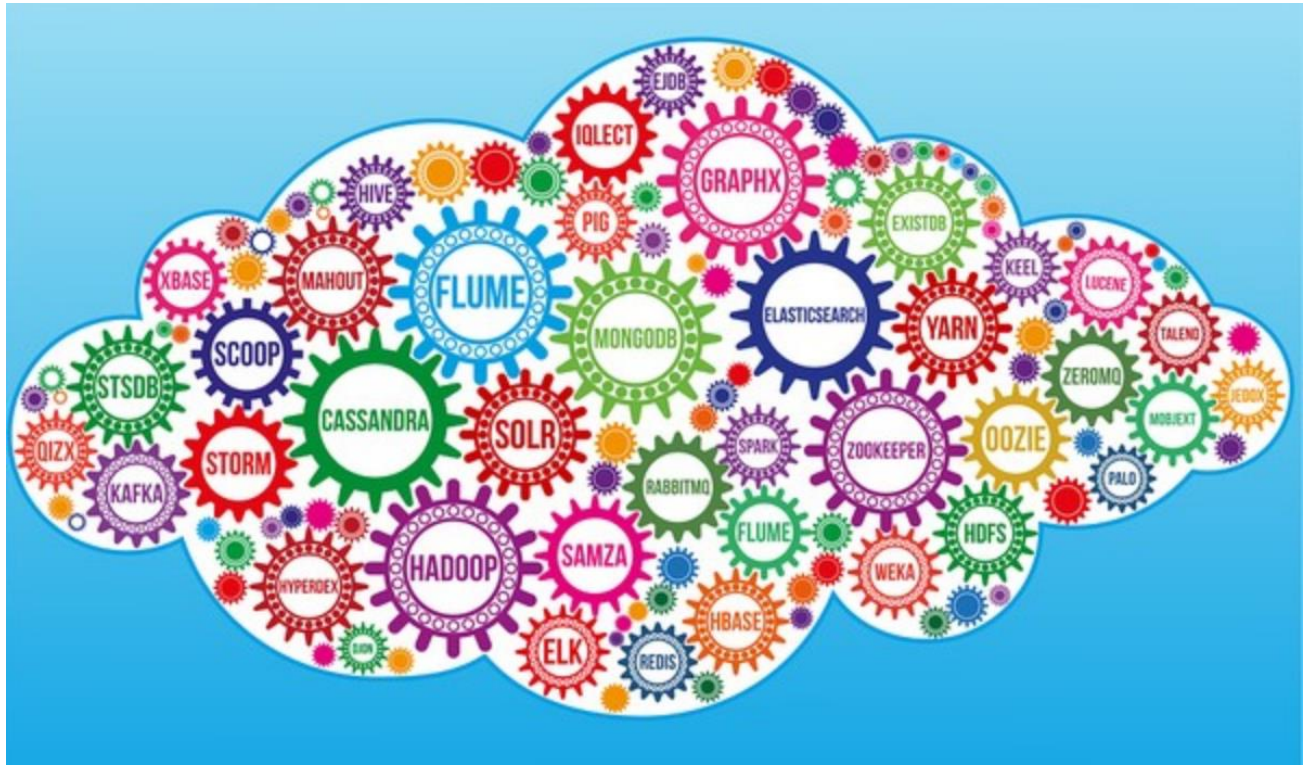


Data analytics – Data discovery



Bigdata Stack

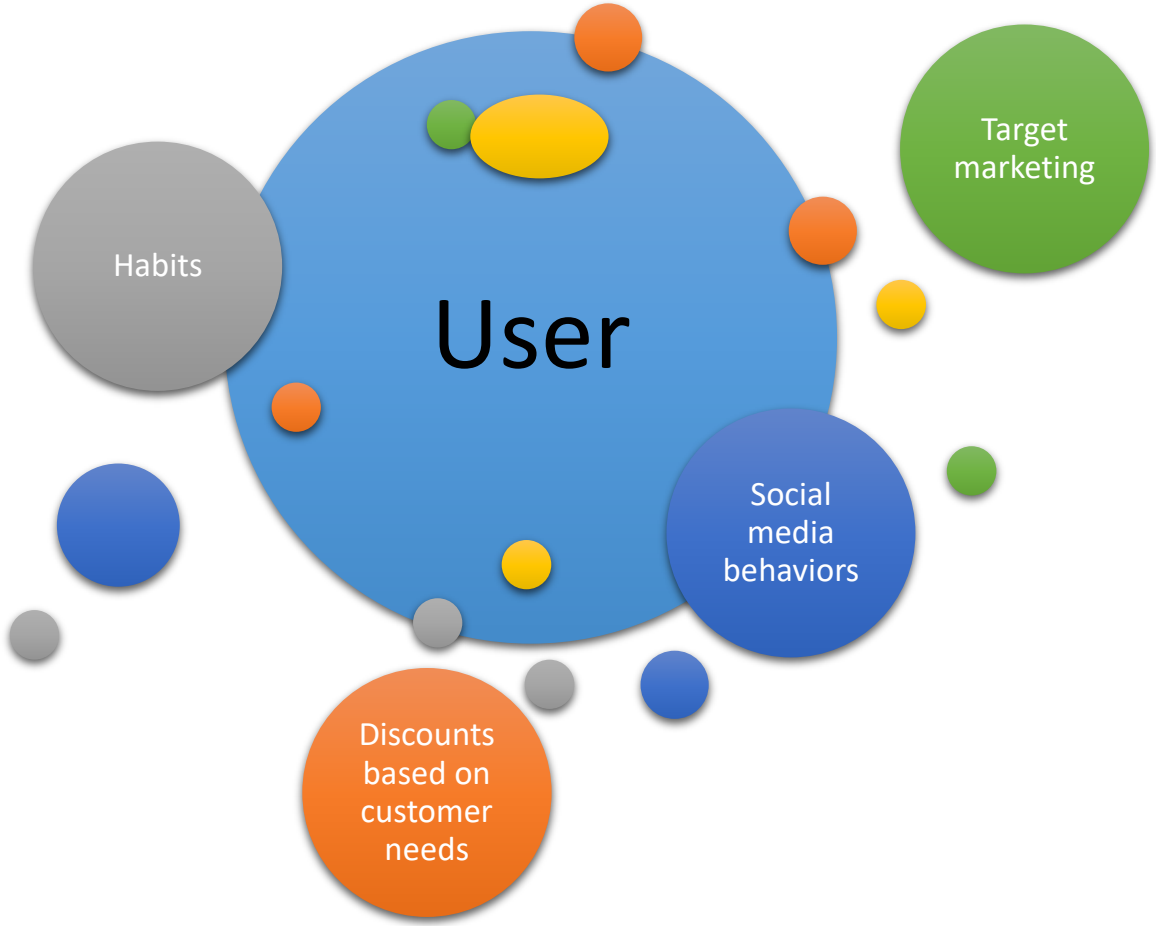




Big Data use cases

- Predictive analytics
- Machine learning (Mllib)
- Anomalies detection
- Fraud detection
- Retail customer 360 customer view
- Scalable and cost effective storage
- Health Care and Genomics

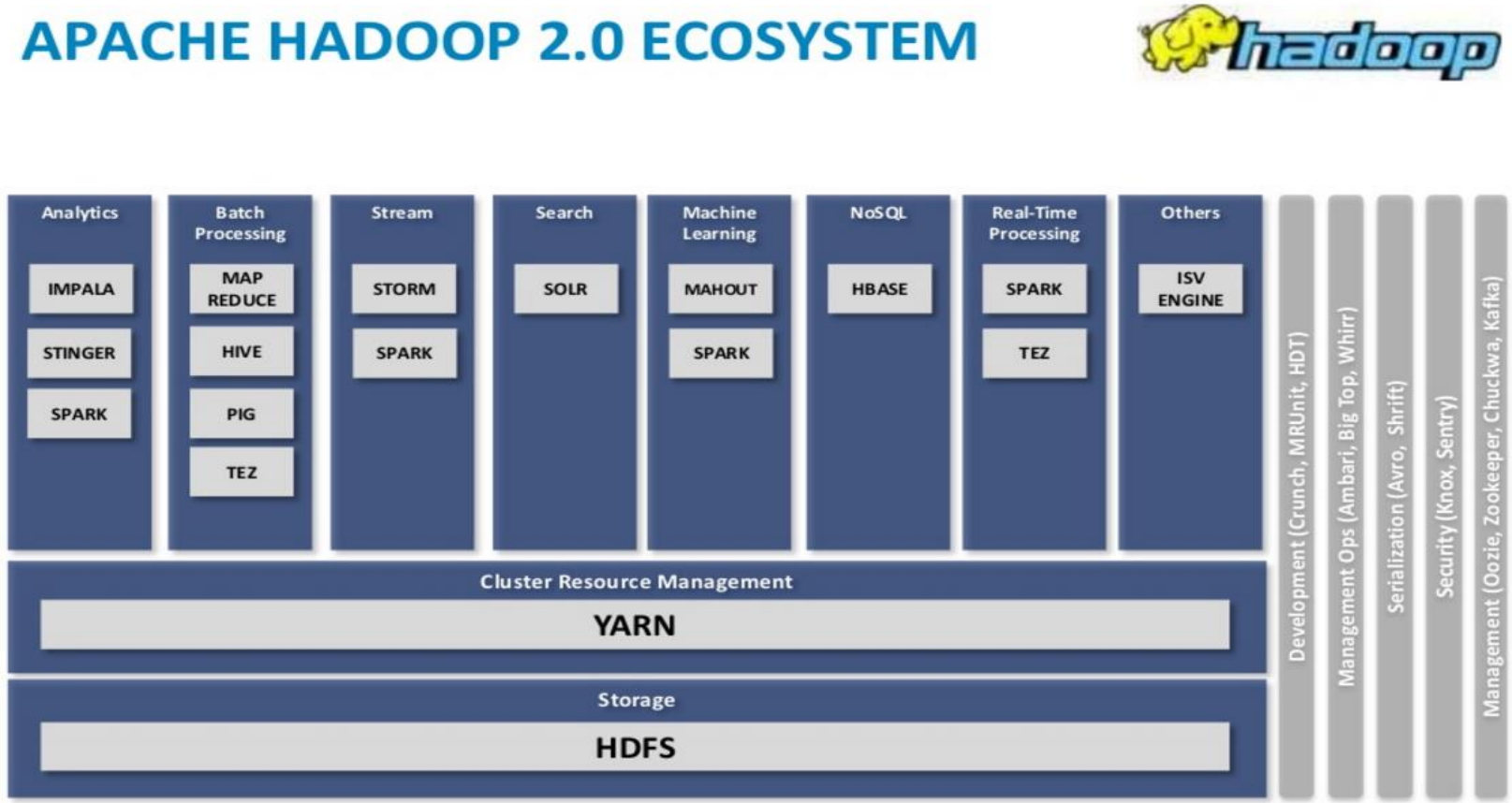
Retail customer 360 customer view



Data Engineer Vs Data Scientist

	Data Engineer	Data Scientist
Role	To engineer software solutions.	To solve business problems using data.
Skills	More of programing and technical skills and ability to architect technical solutions.	Strong of Mathematical Skills and understanding of statistical Models.

Hadoop 2.0 Ecosystem ..cont



http://incubator.apache.org/projects/

Toolkit we will use

- Twitter application API for data ingestion
- Nifi – streaming packet delivery system to capture tweets
- Hive/Impala – Hadoop Query engine
- R – data discovery, filter dataset
- Spark – In Memory data processing
  - Spark via Scala
  - Spark via PySpark
- Solr – Data discovery and visualization

Git repo

- Script and config files will be checked into git repository
- Git repo = https://github.com/abdul-git/hadoop-advance
- Install git on Hadoop nodes: sudo yum -y install git
- To clone git: git clone https://github.com/abdul-git/hadoop-advance