edureka!

Hadoop Ecosystem

Agenda for today's Session

edureka!

Hadoop Ecosystem

- HDFS -> Hadoop Distributed File System
- YARN -> Yet Another Resource Negotiator
- MapReduce -> Data processing using programming
- Spark -> In-memory Data Processing
- PIG, HIVE-> Data Processing Services using Query (SQL-like)
- HBase -> NoSQL Database
- Mahout, Spark MLlib -> Machine Learning
- Apache Drill -> SQL on Hadoop
- Zookeeper -> Managing Cluster
- Oozie -> Job Scheduling
- Flume, Sqoop -> Data Ingesting Services
- Solr & Lucene -> Searching & Indexing
- Ambari -> Provision, Monitor and Maintain cluster

Agenda for today's Session

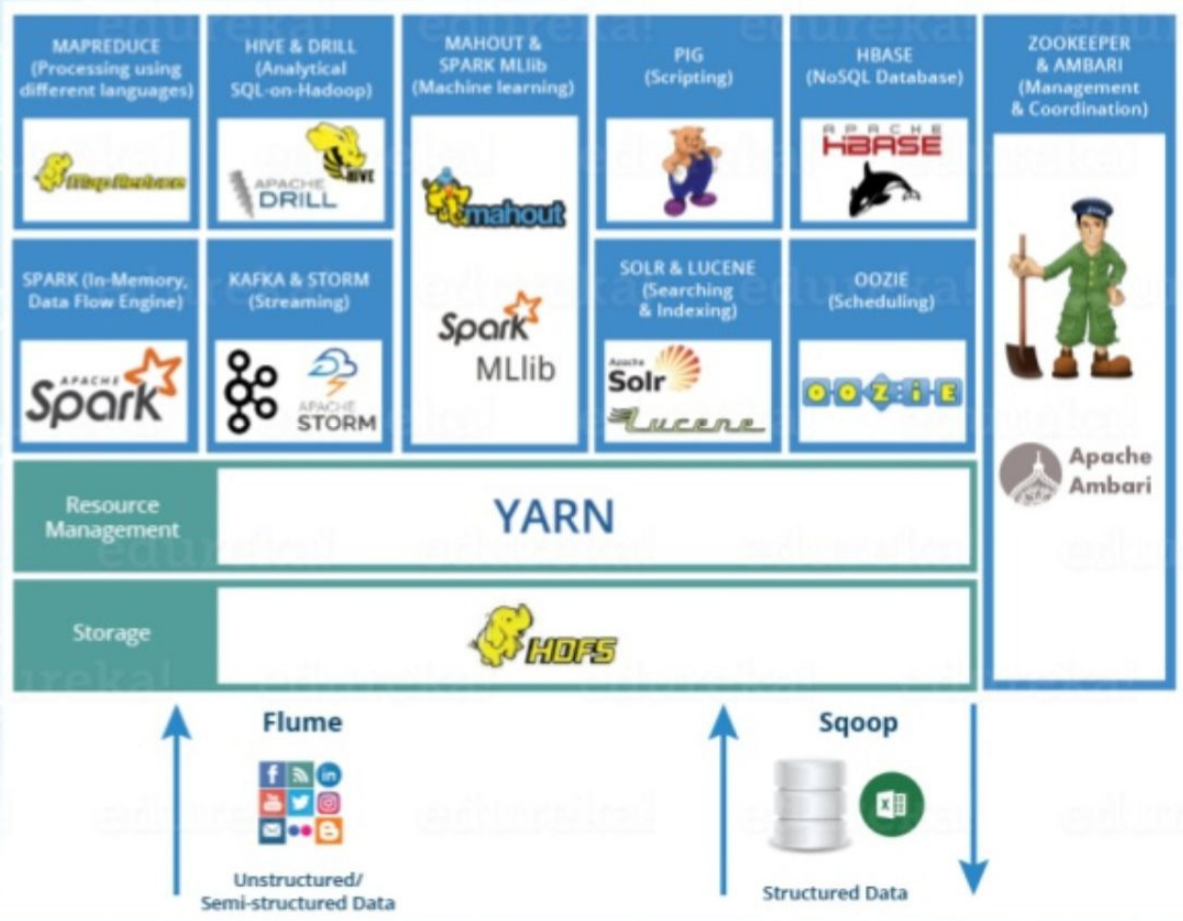
edureka!

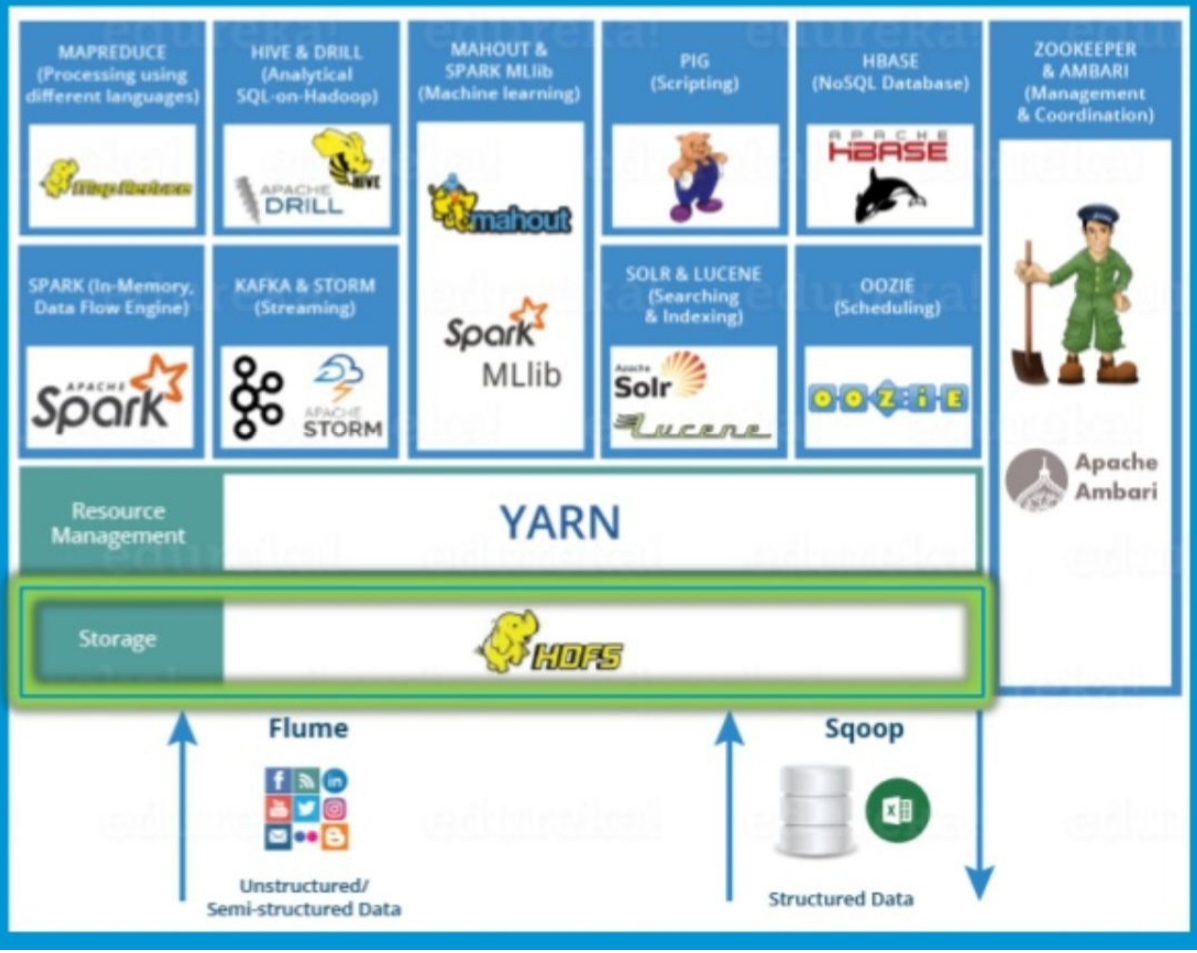
Hadoop Ecosystem

- HDFS -> Hadoop Distributed File System
- YARN -> Yet Another Resource Negotiator
- MapReduce -> Data processing using programming
- Spark -> In-memory Data Processing
- PIG, HIVE-> Data Processing Services using Query (SQL-like)
- HBase -> NoSQL Database
- Mahout, Spark MLlib -> Machine Learning
- Apache Drill -> SQL on Hadoop
- Zookeeper -> Managing Cluster
- Oozie -> Job Scheduling
- Flume, Sqoop -> Data Ingesting Services
- Solr & Lucene -> Searching & Indexing
- Ambari -> Provision, Monitor and Maintain cluster

Hadoop Ecosystem

edureka!

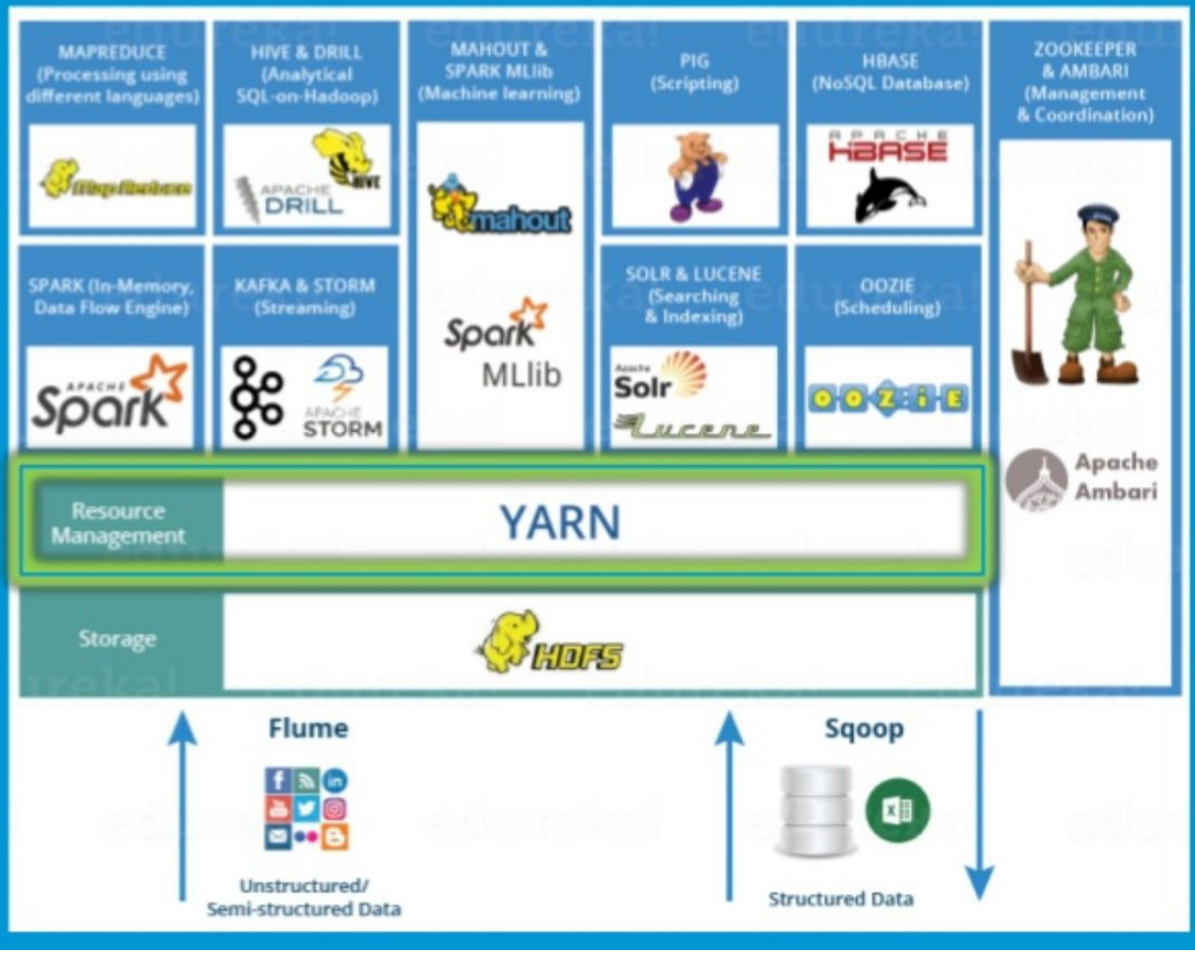



MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Semi-structured Data
Sqoop
Structured Data

HDFS (Hadoop Distributed File System) - Storage

edureka!

- Stores different types of large data sets (i.e. structured, unstructured and semi structured data)
- HDFS creates a level of abstraction over the resources, from where we can see the whole HDFS as a single unit
- Stores data across various nodes and maintains the log file about the stored data (metadata)
- HDFS has two core components, i.e. NameNode and DataNode


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Semi-structured Data
Sqoop
Structured Data

YARN (Yet Another Resource Negotiator)

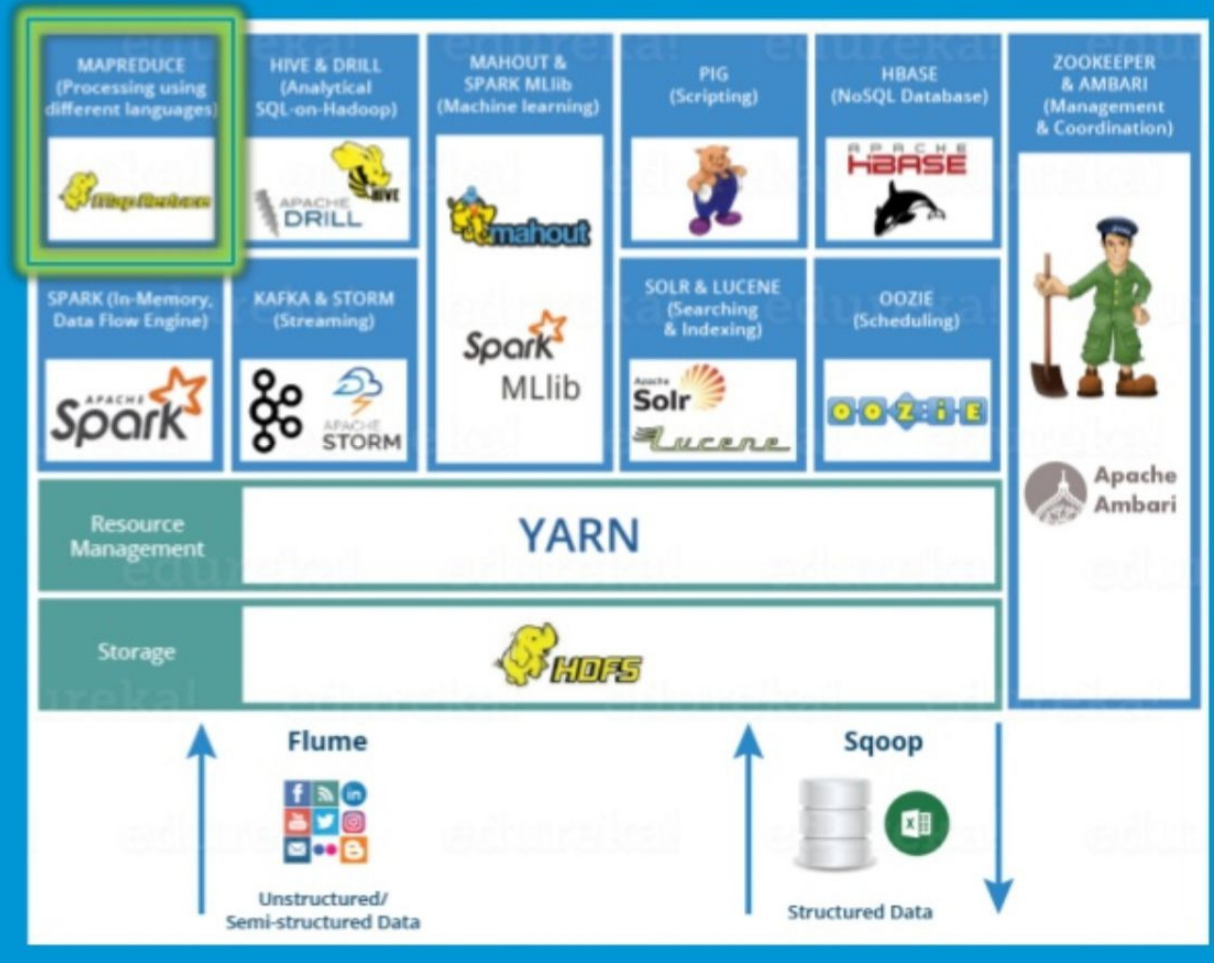
edureka!

- Performs all your processing activities by allocating resources and scheduling tasks
- Two services: ResourceManager and NodeManager
- ResourceManager: Manages resources and schedule applications running on top of YARN
- NodeManager: Manages containers and monitors resource utilization in each container

YARN (Yet Another Resource Negotiator)

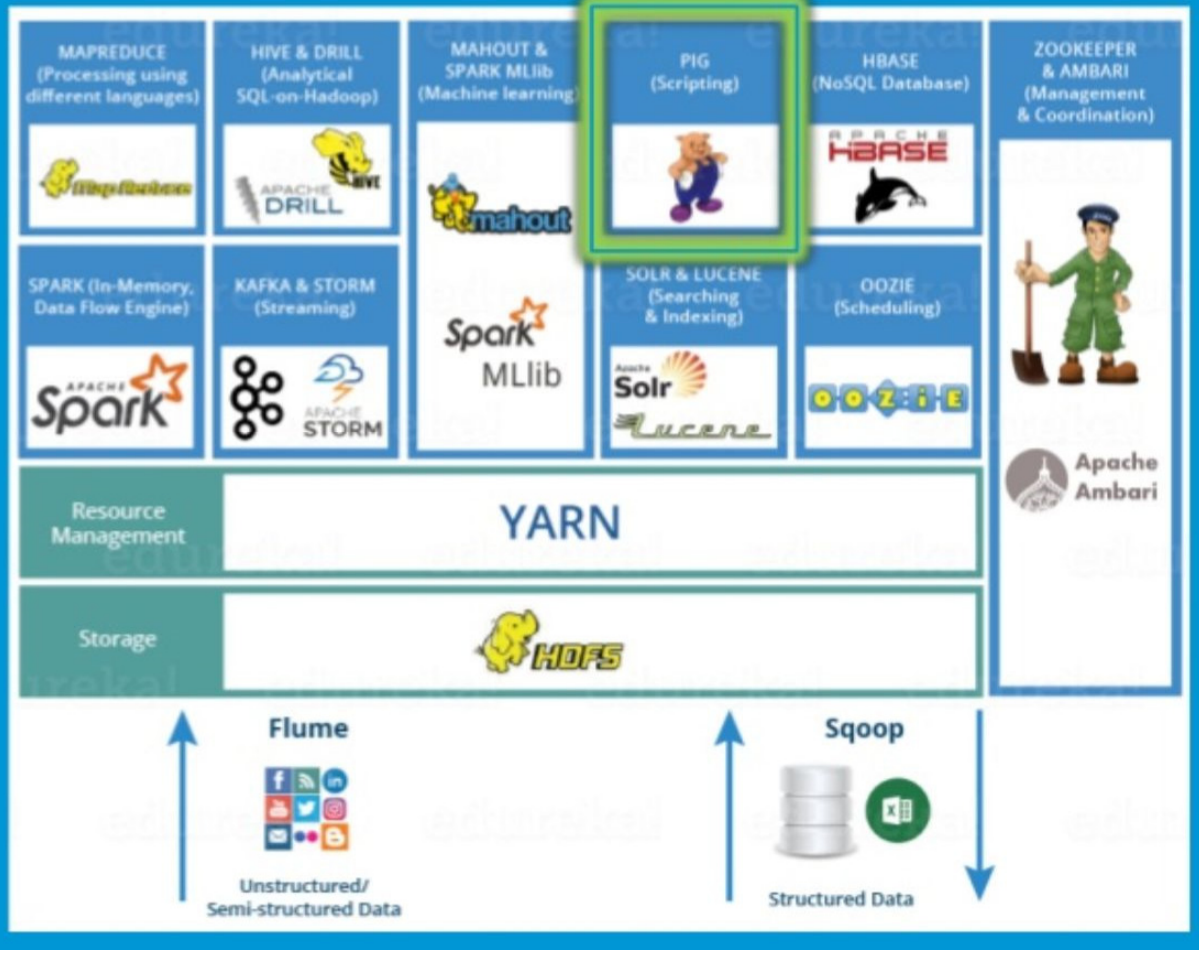
edureka!

- Performs all your processing activities by allocating resources and scheduling tasks
- Two services: ResourceManager and NodeManager
- ResourceManager: Manages resources and schedule applications running on top of YARN
- NodeManager: Manages containers and monitors resource utilization in each container


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Semi-structured Data
Sqoop
Structured Data

MapReduce: Data Processing Using Programming edureka!

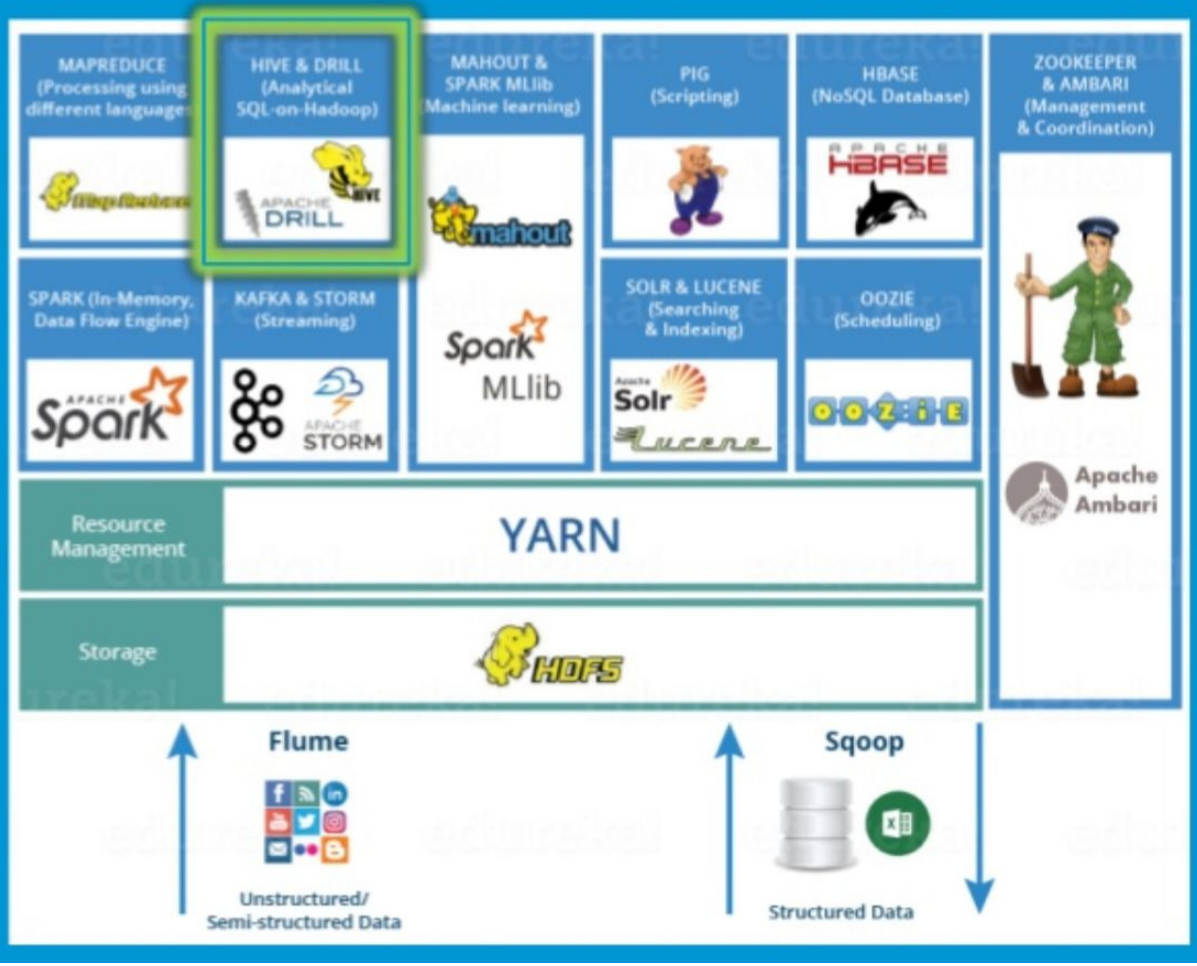
- Core component in a Hadoop Ecosystem for processing
- Helps in writing applications that processes large data sets using distributed and parallel algorithms
- In a MapReduce program, Map() and Reduce() are two functions
- Map function performs actions like filtering, grouping and sorting
- Reduce function aggregates and summarizes the result produced by map function


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Semi-structured Data
Sqoop
Structured Data

PIG: Data Processing Service using Query

edureka!

- PIG has two parts: Pig Latin, the language and the pig runtime, for the execution environment
- **1 line of pig latin = approx. 100 lines of Map-Reduce job**
- The compiler internally converts pig latin to MapReduce
- It gives you a platform for building data flow for ETL (Extract, Transform and Load)
- PIG first loads the data, then performs various functions like grouping, filtering, joining, sorting, etc. and finally dumps the data on the screen or stores in HDFS.


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Semi-structured Data
Sqoop
Structured Data

HIVE: Data Processing Service using Query

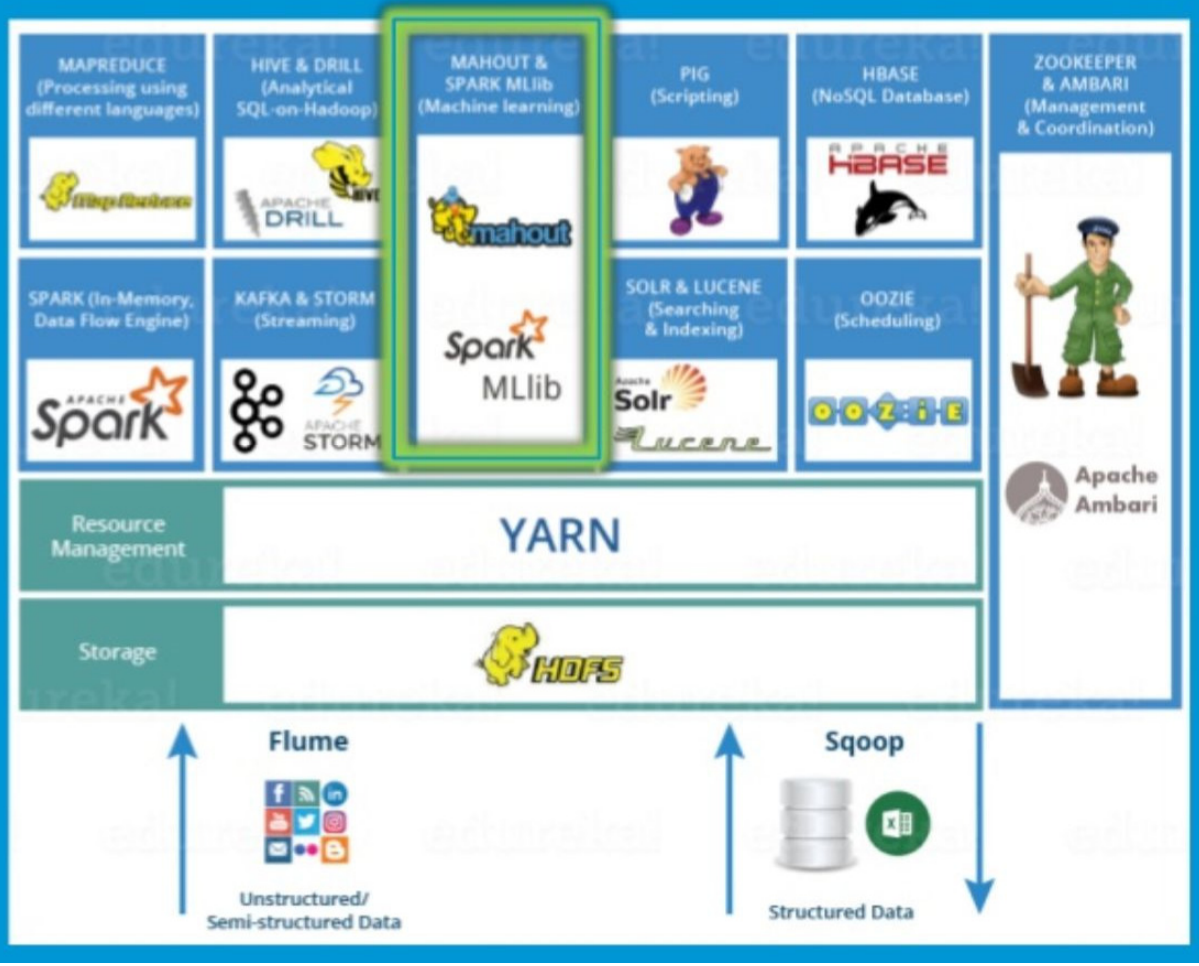
edureka!

- A data warehousing component which analyses data sets in a distributed environment using SQL-like interface
- The query language of Hive is called Hive Query Language(HQL)
- 2 basic components: Hive Command Line and JDBC/ODBC driver
- Supports user defined functions (UDF) to accomplish specific needs

HIVE: Data Processing Service using Query

edureka!

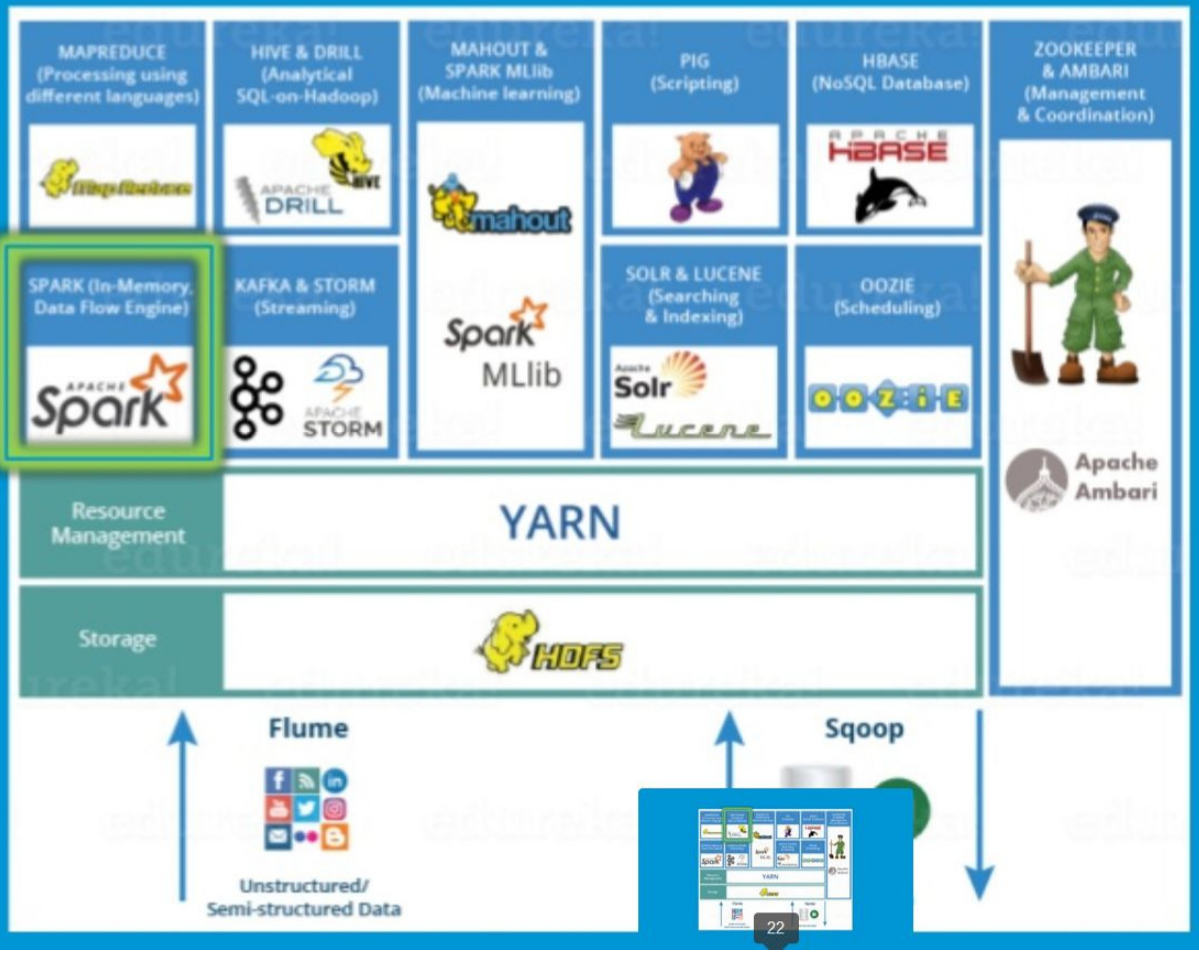
- A data warehousing component which analyses data sets in a distributed environment using SQL-like interface
- The query language of Hive is called Hive Query Language(HQL)
- 2 basic components: Hive Command Line and JDBC/ODBC driver
- Supports user defined functions (UDF) to accomplish specific needs


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Semi-structured Data
Sqoop
Structured Data

Mahout: Machine Learning

edureka!

- Provides an environment for creating machine learning applications
- It performs collaborative filtering, clustering and classification
- Provides a command line to invoke various algorithms.
- It has a predefined set of library which already contains different inbuilt algorithms for different use cases.


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Sqoop
Unstructured/
Semi-structured Data

Spark: In-memory Data Processing

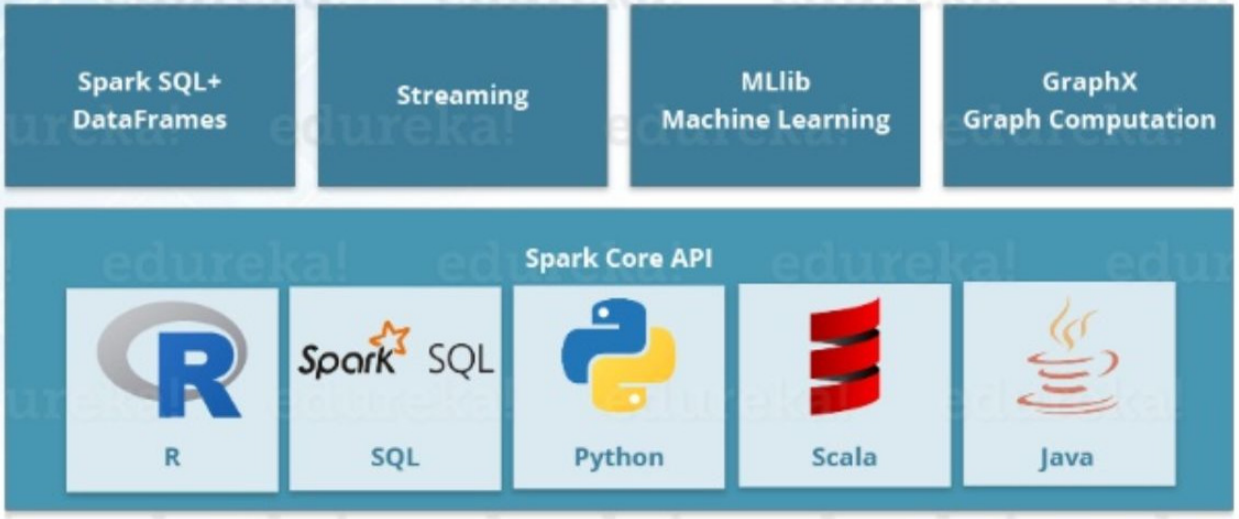
edureka!

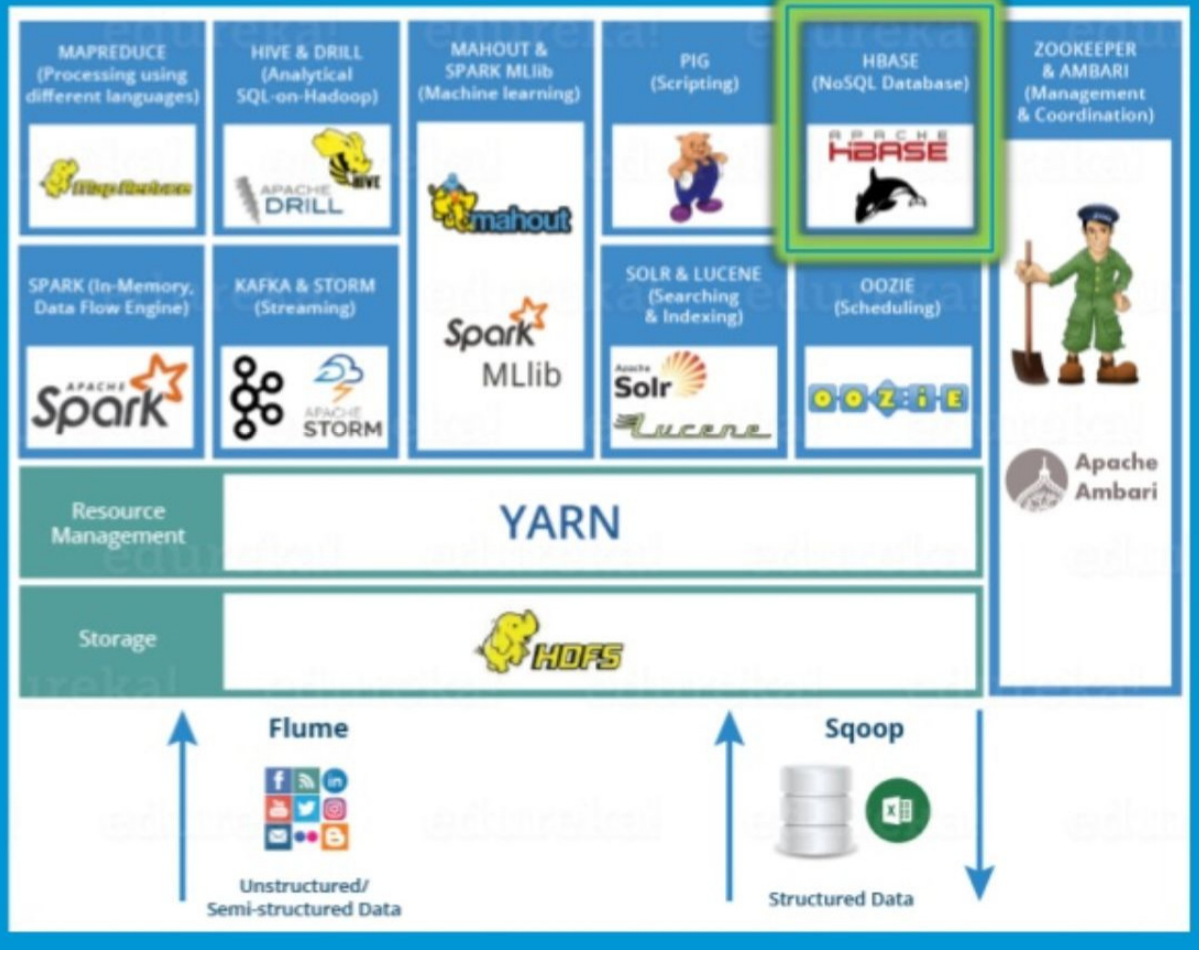
- A framework for real time data analytics in a distributed computing environment.
- Written in Scala and was originally developed at the University of California, Berkeley.
- It executes in-memory computations to increase speed of data processing over Map-Reduce.
- 100x faster than Hadoop for large scale data processing by exploiting in-memory computations

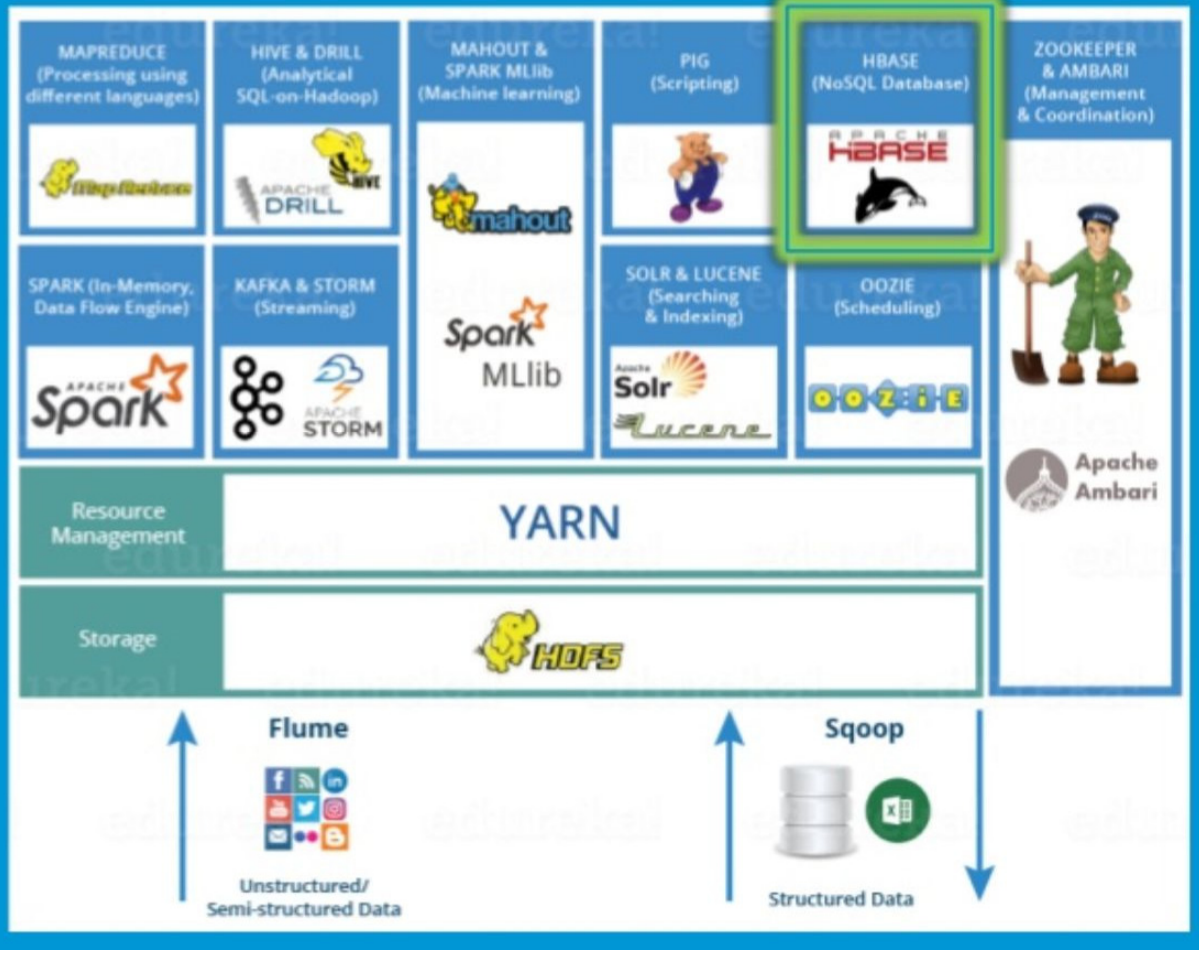
Spark: In-memory Data Processing

edureka!

- Spark comes packed with high-level libraries
- Provides various services like MLlib, GraphX, SQL + Data Frames, Streaming services
- Supports various languages like R, SQL, Python, Scala, Java
- Seamlessly integrates in complex workflow



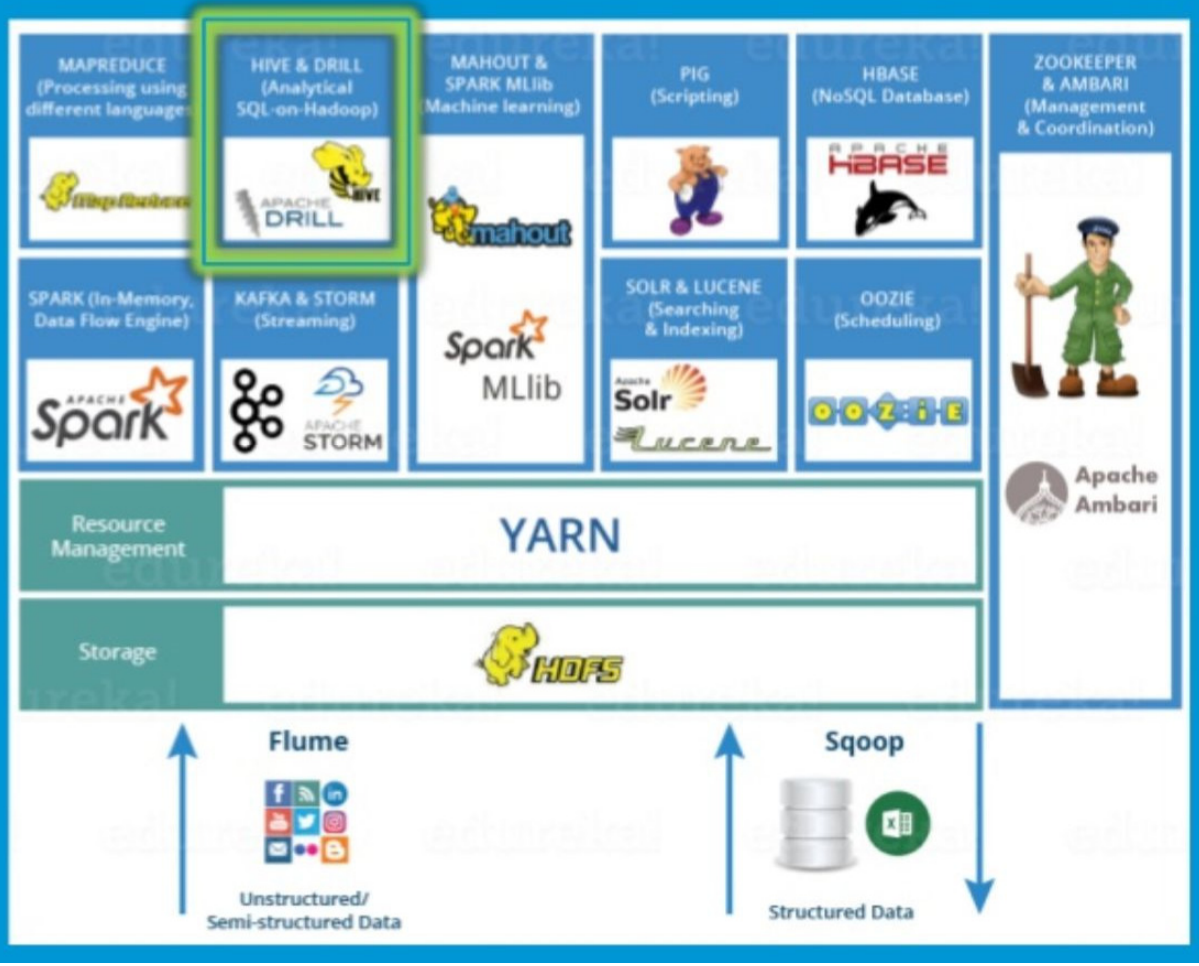

MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Semi-structured Data
Sqoop
Structured Data


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Semi-structured Data
Sqoop
Structured Data

Hbase: NoSQL Database

edureka!

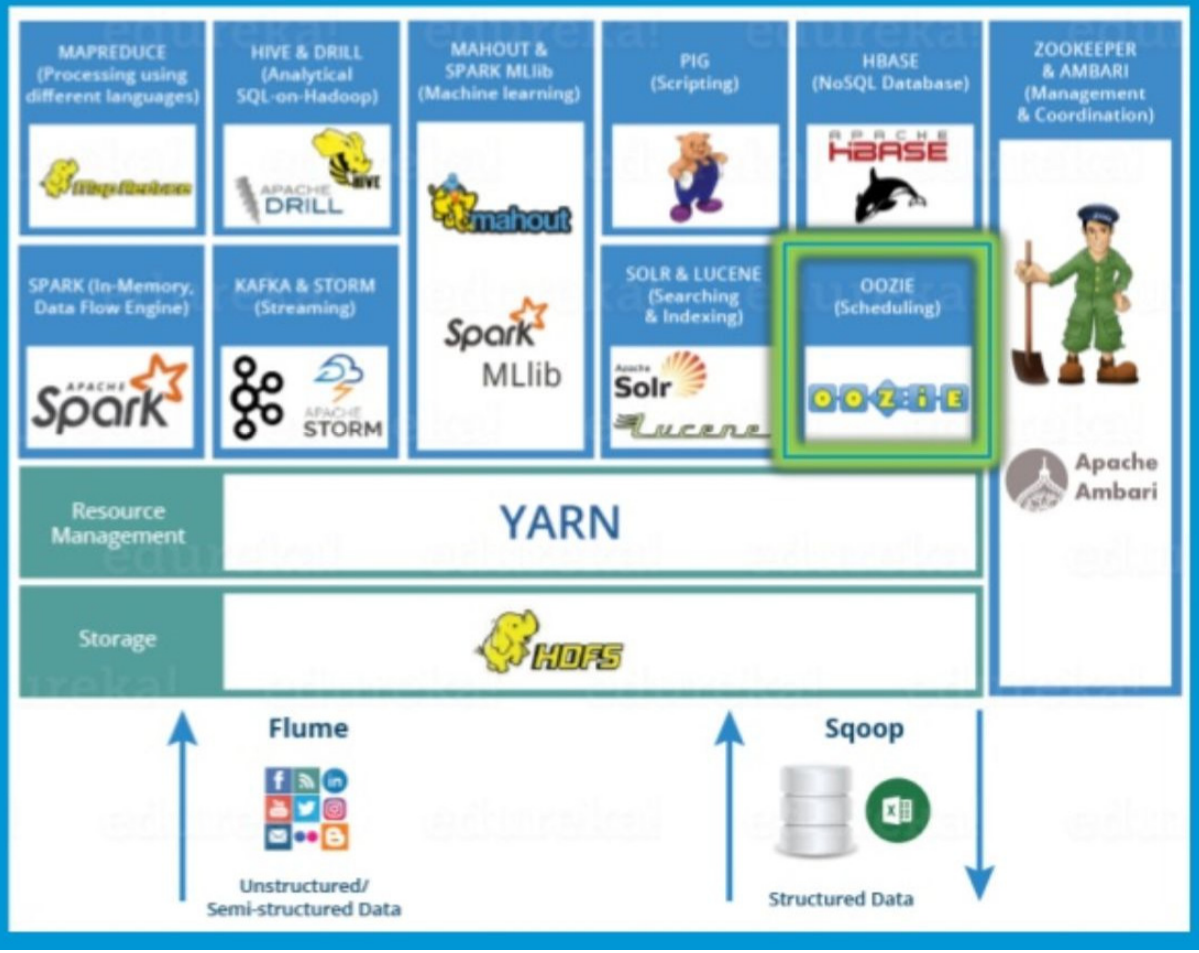
- An open source, non-relational distributed database - a **NoSQL** database
- Supports all types of data and that is why, it's capable of handling anything and everything
- It is modelled after Google's BigTable
- It gives us a fault tolerant way of storing sparse data
- It is written in Java, and HBase applications can be written in REST, Avro and Thrift APIs


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
APACHE DRILL
MAHOUT & SPARK MLlib
(Machine learning)
mahout
Spark MLlib
PIG
(Scripting)
HBASE
(NoSQL Database)
APACHE HBASE
ZOOKEEPER & AMBARI
(Management & Coordination)
Apache Ambari
SPARK (In-Memory, Data Flow Engine)
APACHE Spark
KAFKA & STORM
(Streaming)
APACHE STORM
SOLR & LUCENE
(Searching & Indexing)
Apache Solr
lucene
OOZIE
(Scheduling)
OOZIE
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data

Apache Drill: SQL on Hadoop

edureka!

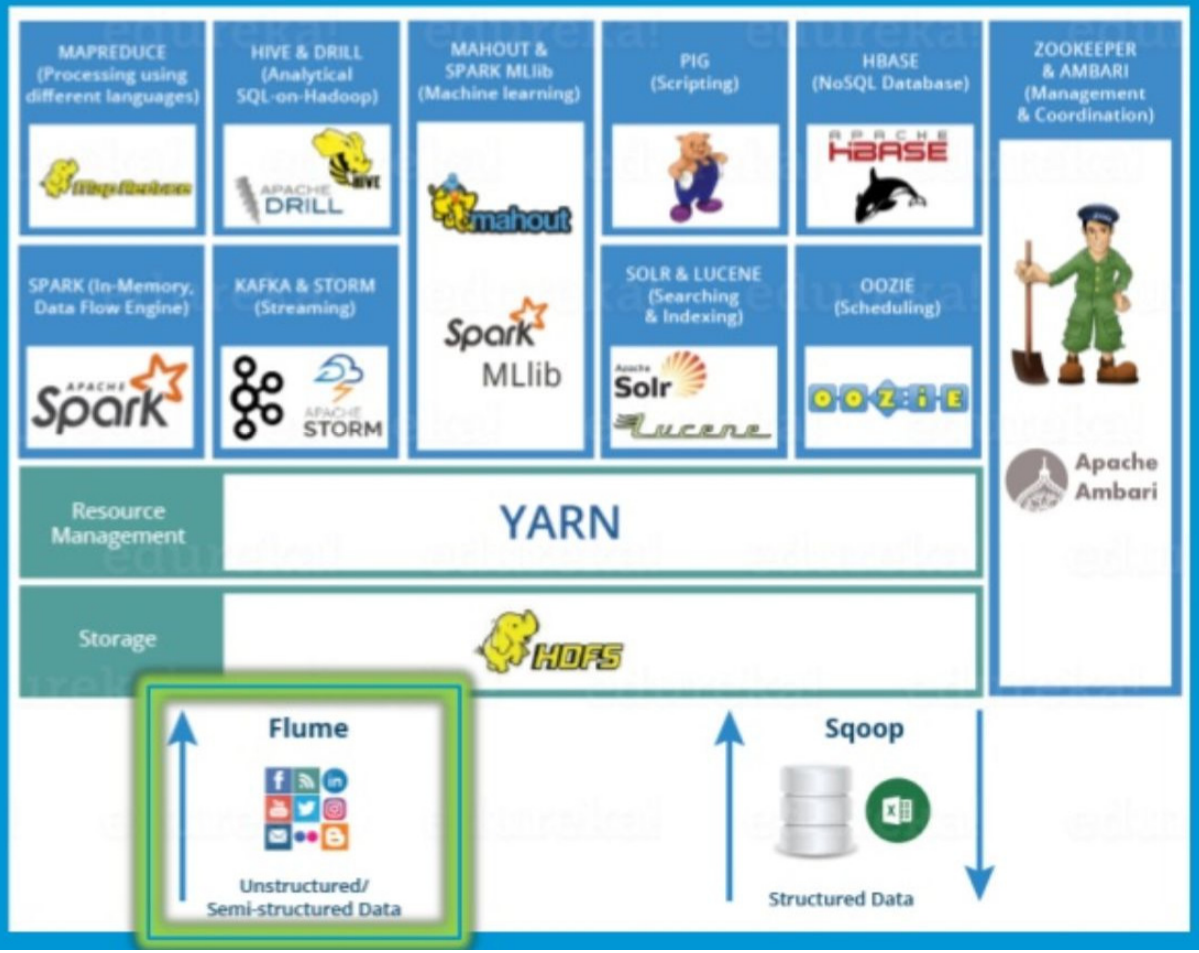
- An open source application which works with distributed environment to analyze large data sets
- Follows the ANSI SQL
- Supports different kinds NoSQL databases and file systems
- For example: Azure Blob Storage, Google Cloud Storage, HBase, MongoDB, MapR-DB HDFS, MapR-FS, Amazon S3, Swift, NAS and local files
- Combines a variety of data stores just by using a single query


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Semi-structured Data
Sqoop
Structured Data

Oozie: Job Scheduler

edureka!

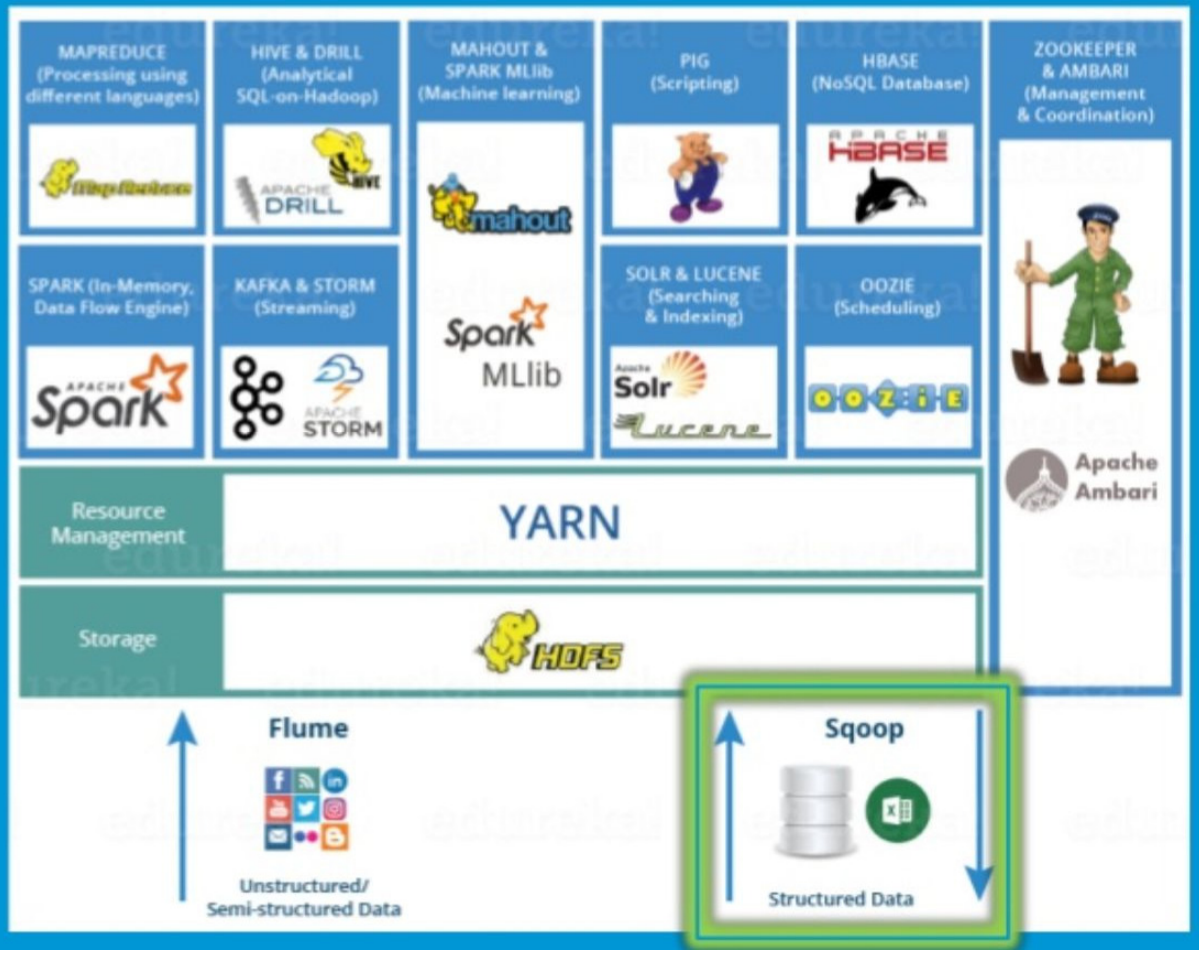
- Oozie is a job scheduler in Hadoop ecosystem
- Two kinds of Oozie jobs: Oozie workflow and Oozie Coordinator
- **Oozie workflow**: Sequential set of actions to be executed
- **Oozie Coordinator**: Oozie jobs which are triggered when the data is made available to it or even triggered based on time


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Semi-structured Data
Sqoop
Structured Data

Apache Flume: Data Ingesting Service

edureka!

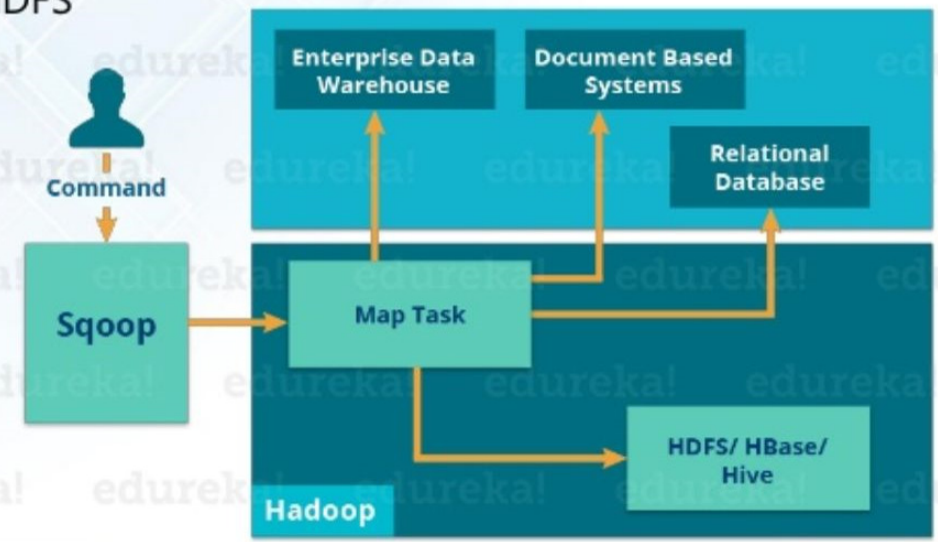
- Ingests unstructured and semi-structured data into HDFS.
- It helps us in collecting, aggregating and moving large amount of data sets.
- It helps us to ingest online streaming data from various sources like network traffic, social media, email messages, log files etc. in HDFS.

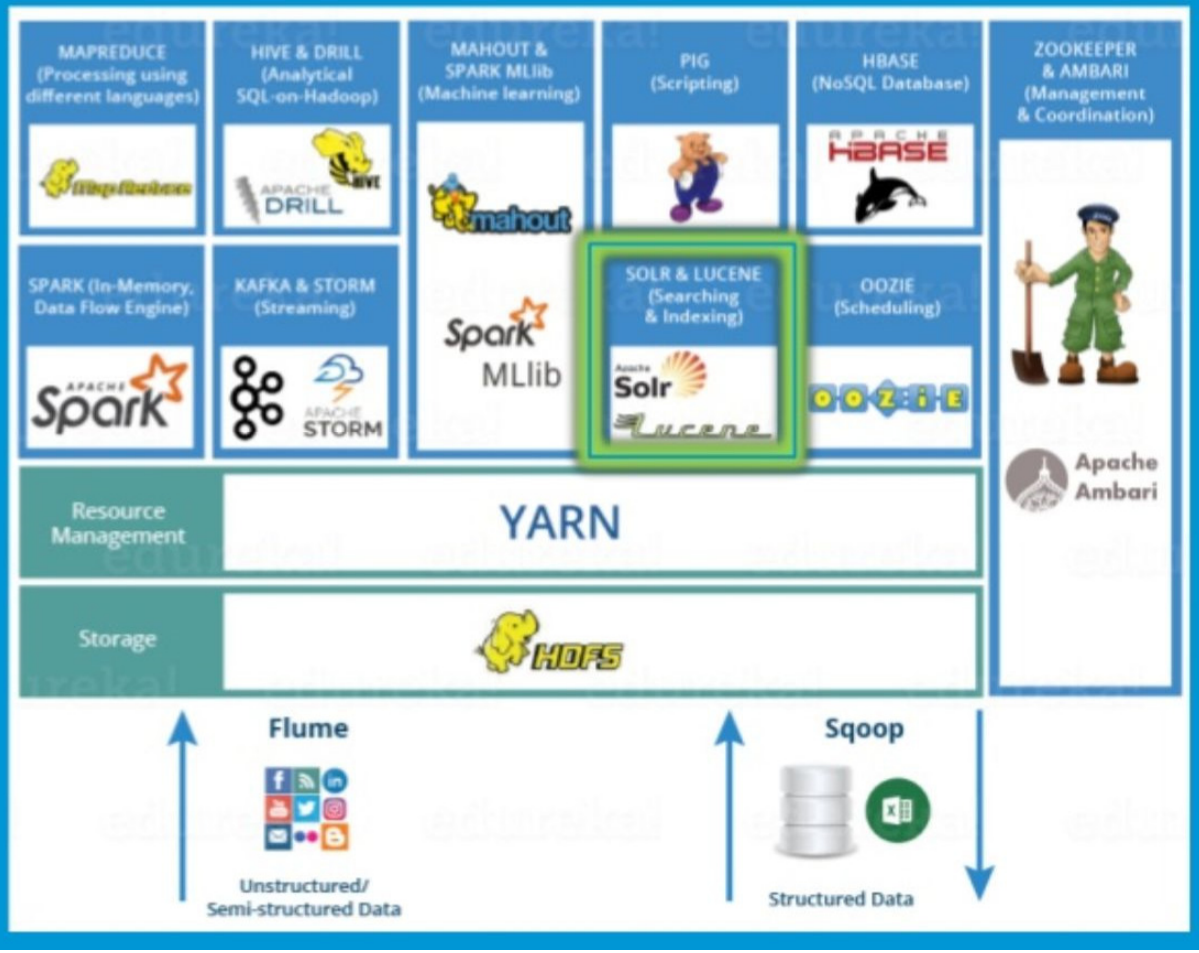

MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
APACHE DRILL
MAHOUT & SPARK MLlib
(Machine learning)
mahout
Spark MLlib
PIG
(Scripting)
HBASE
(NoSQL Database)
APACHE HBASE
ZOOKEEPER & AMBARI
(Management & Coordination)
Apache Ambari
SPARK (In-Memory, Data Flow Engine)
APACHE Spark
KAFKA & STORM
(Streaming)
APACHE STORM
SOLR & LUCENE
(Searching & Indexing)
Apache Solr
lucene
OOZIE
(Scheduling)
OOZIE
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Semi-structured Data
Sqoop
Structured Data

Apache Sqoop: Data Ingesting Service

edureka!

- Another data ingesting service
- Sqoop can import as well as export structured data from RDBMS
- Flume only ingests unstructured data or semi-structured data into HDFS

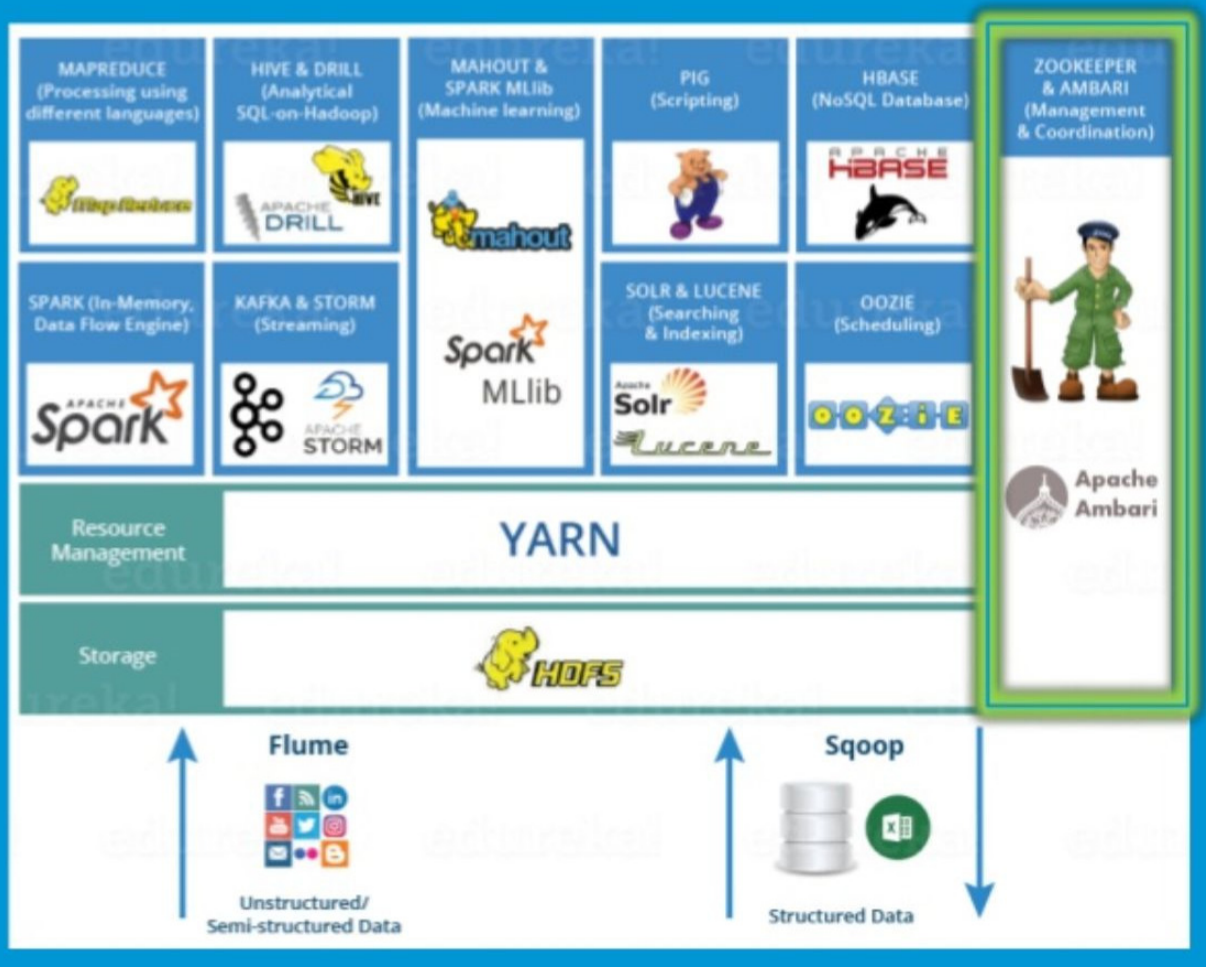



MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data

Apache Solr and Lucene

edureka!

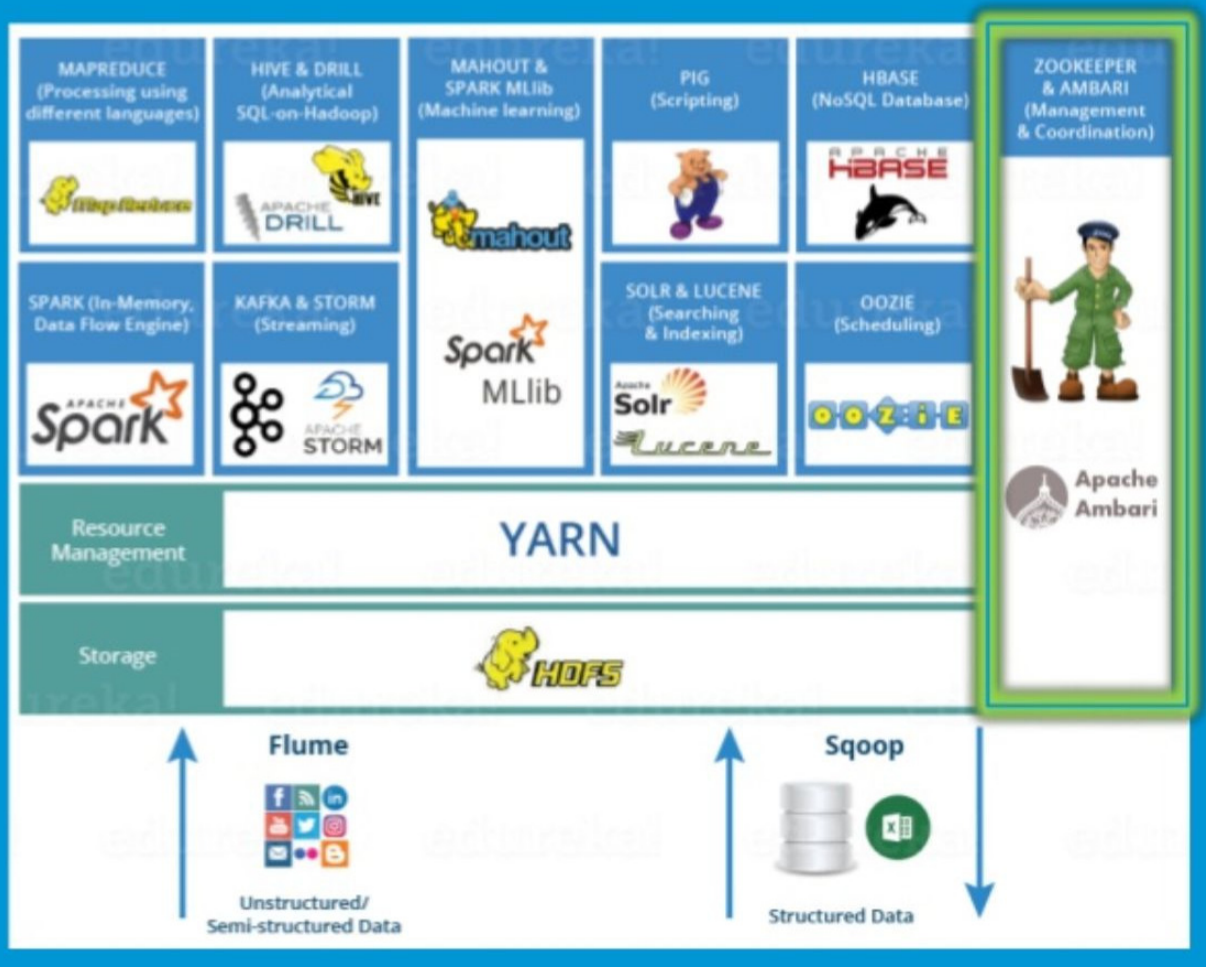
- Two services which are used for searching and indexing in Hadoop Ecosystem
- Apache Lucene is based on Java, which also helps in spell checking
- Apache Lucene is the engine, Apache Solr is a complete application built around Lucene
- Solr uses Apcahe Lucene Java search library for searching and indexing


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
APACHE DRILL
MAHOUT & SPARK MLlib
(Machine learning)
mahout
Spark MLlib
PIG
(Scripting)
HBASE
(NoSQL Database)
APACHE HBASE
ZOOKEEPER & AMBARI
(Management & Coordination)
Apache Ambari
SPARK (In-Memory, Data Flow Engine)
APACHE Spark
KAFKA & STORM
(Streaming)
APACHE STORM
SOLR & LUCENE
(Searching & Indexing)
Apache Solr
lucene
OOZIE
(Scheduling)
OOZIE
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data

ZooKeeper: Coordinator

edureka!

- An open-source server which enables highly reliable distributed coordination
- Apache Zookeeper coordinates with various Hadoop services in a distributed environment
- Performs synchronization, configuration maintenance, grouping and naming


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
APACHE DRILL
MAHOUT & SPARK MLlib
(Machine learning)
mahout
Spark MLlib
PIG
(Scripting)
HBASE
(NoSQL Database)
APACHE HBASE
ZOOKEEPER & AMBARI
(Management & Coordination)
Apache Ambari
SPARK (In-Memory, Data Flow Engine)
APACHE Spark
KAFKA & STORM
(Streaming)
APACHE STORM
SOLR & LUCENE
(Searching & Indexing)
Apache Solr
lucene
OOZIE
(Scheduling)
OOZIE
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data

Apache Ambari: Cluster Manager

edureka!

- Software for provisioning, managing and monitoring Apache Hadoop clusters
- Gives us step by step process for installing Hadoop services
- Handles configuration of Hadoop services
- Provides a central management service for starting, stopping and re-configuring Hadoop services
- Monitors health and status of the Hadoop cluster

Apache Ambari: Cluster Manager

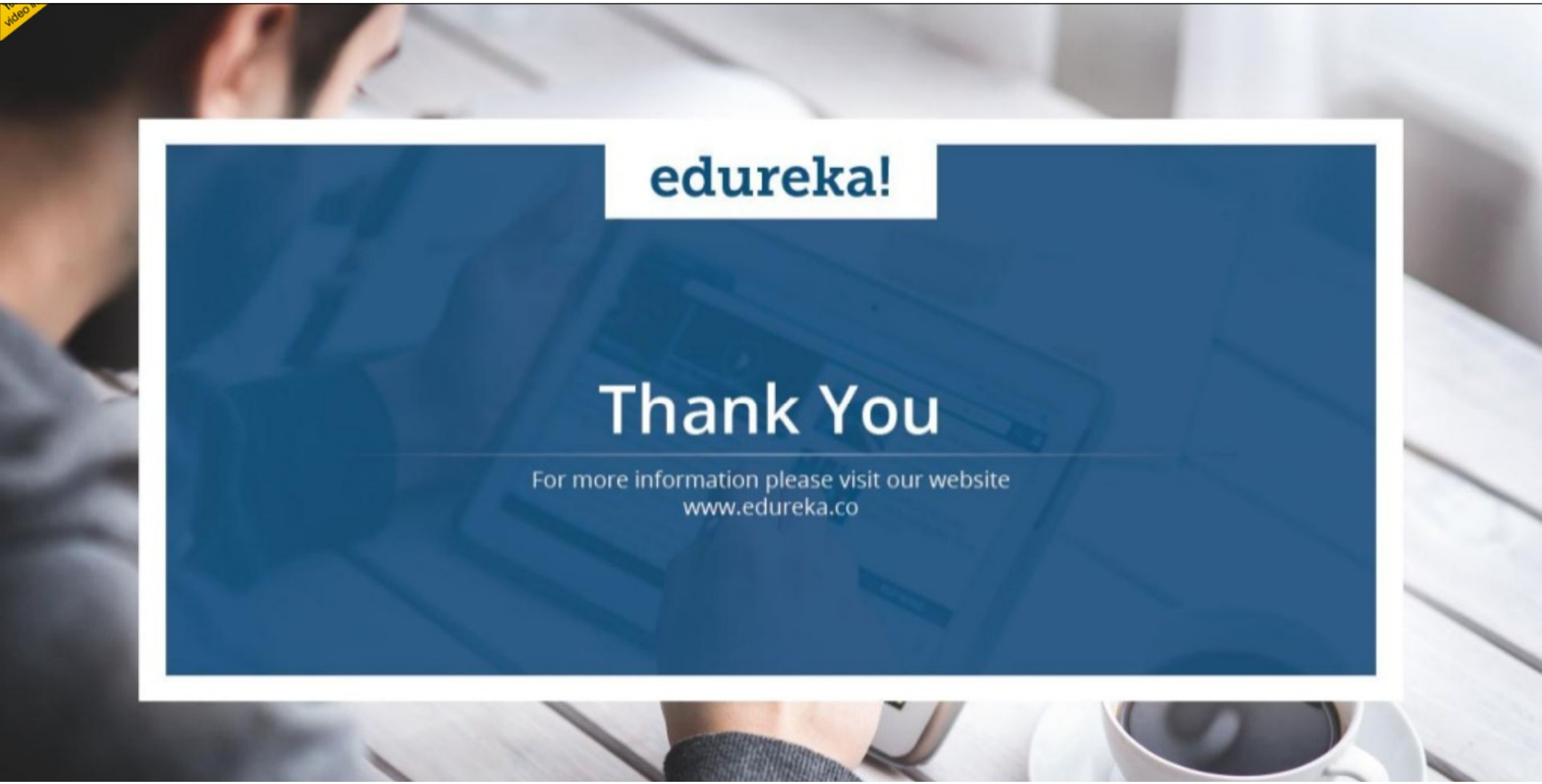
edureka!

- Software for provisioning, managing and monitoring Apache Hadoop clusters
- Gives us step by step process for installing Hadoop services
- Handles configuration of Hadoop services
- Provides a central management service for starting, stopping and re-configuring Hadoop services
- Monitors health and status of the Hadoop cluster

Learning Resources

edureka!

- Hadoop Tutorial: www.edureka.co/blog/hadoop-tutorial
- HDFS Ecosystem: www.edureka.co/blog/hadoop-ecosystem


edureka!
Thank You
For more information please visit our website
www.edureka.co