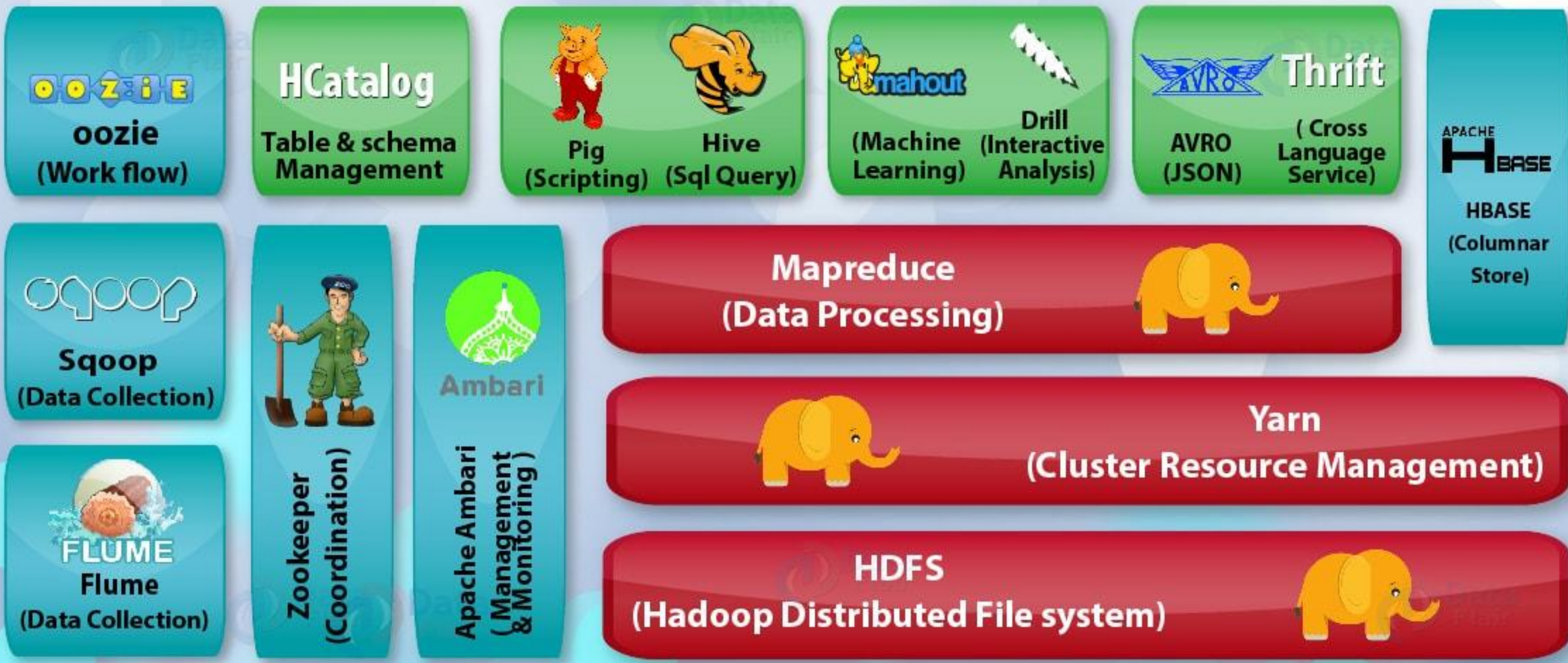
Hadoop Ecosystem

Overview of Hadoop Ecosystem.

Hadoop Ecosystem



oozie
(Work flow)

HCatalog
Table & schema Management

Pig
(Scripting)

Hive
(Sql Query)

mahout
(Machine Learning)

Drill
(Interactive Analysis)

AVRO
(JSON)

Thrift
(Cross Language Service)

HBASE
(Columnar Store)

Sqoop
(Data Collection)

Flume
(Data Collection)

Zookeeper
(Coordination)

Ambari

Apache Ambari
(Management & Monitoring)

Mapreduce
(Data Processing)

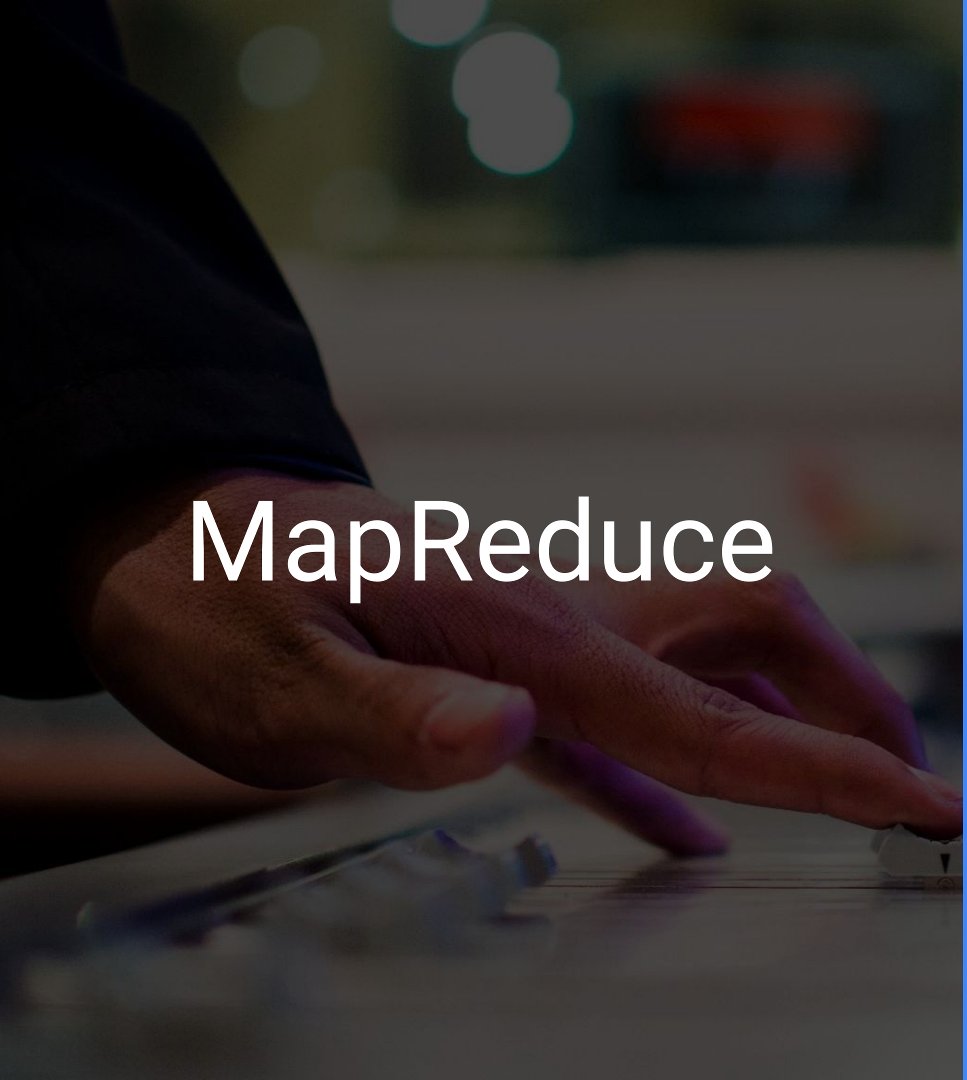
Yarn
(Cluster Resource Management)

HDFS
(Hadoop Distributed File system)

We are going to learn the following Hadoop components.

HDFS

It is the most important component of Hadoop Ecosystem. HDFS is the primary storage system of Hadoop. Hadoop distributed file system (HDFS) is a java based file system that provides scalable, fault tolerance, reliable and cost efficient data storage for Big data. HDFS is a distributed file system that runs on commodity hardware. HDFS is already configured with default configuration for many installations. Most of the time for large clusters configuration is needed. Hadoop interact directly with HDFS by shell-like commands.



MapReduce

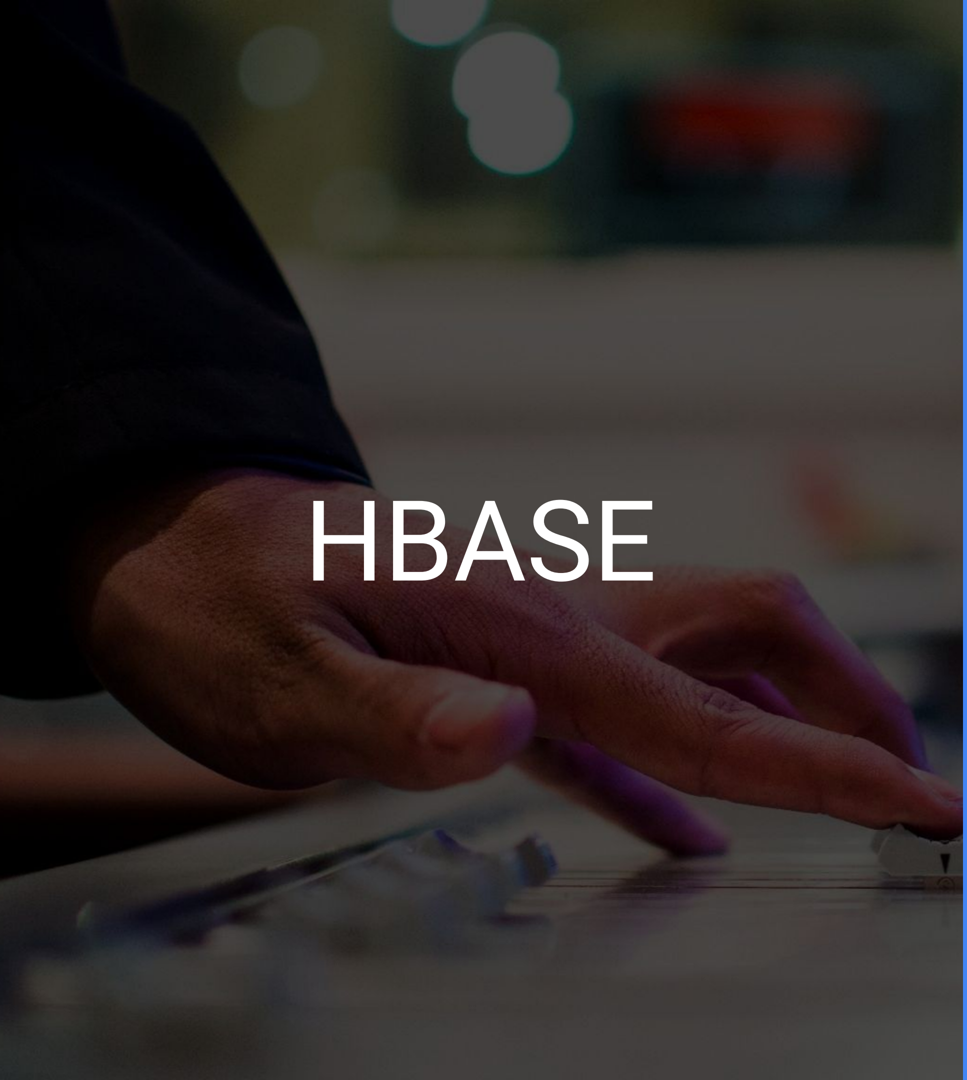
Hadoop MapReduce is the core Hadoop ecosystem component which provides data processing. MapReduce is a software framework for easily writing applications that process the vast amount of structured and unstructured data stored in the Hadoop Distributed File system.

MapReduce programs are parallel in nature, thus are very useful for performing large-scale data analysis using multiple machines in the cluster. Thus, it improves the speed and reliability of cluster this parallel processing.

HIVE

The Hadoop ecosystem component, Apache Hive, is an open source data warehouse system for querying and analyzing large datasets stored in Hadoop files. Hive do three main functions: *data summarization, query, and analysis*.

Hive use language called HiveQL (HQL), which is similar to SQL. HiveQL automatically translates SQL-like queries into MapReduce jobs which will execute on Hadoop.

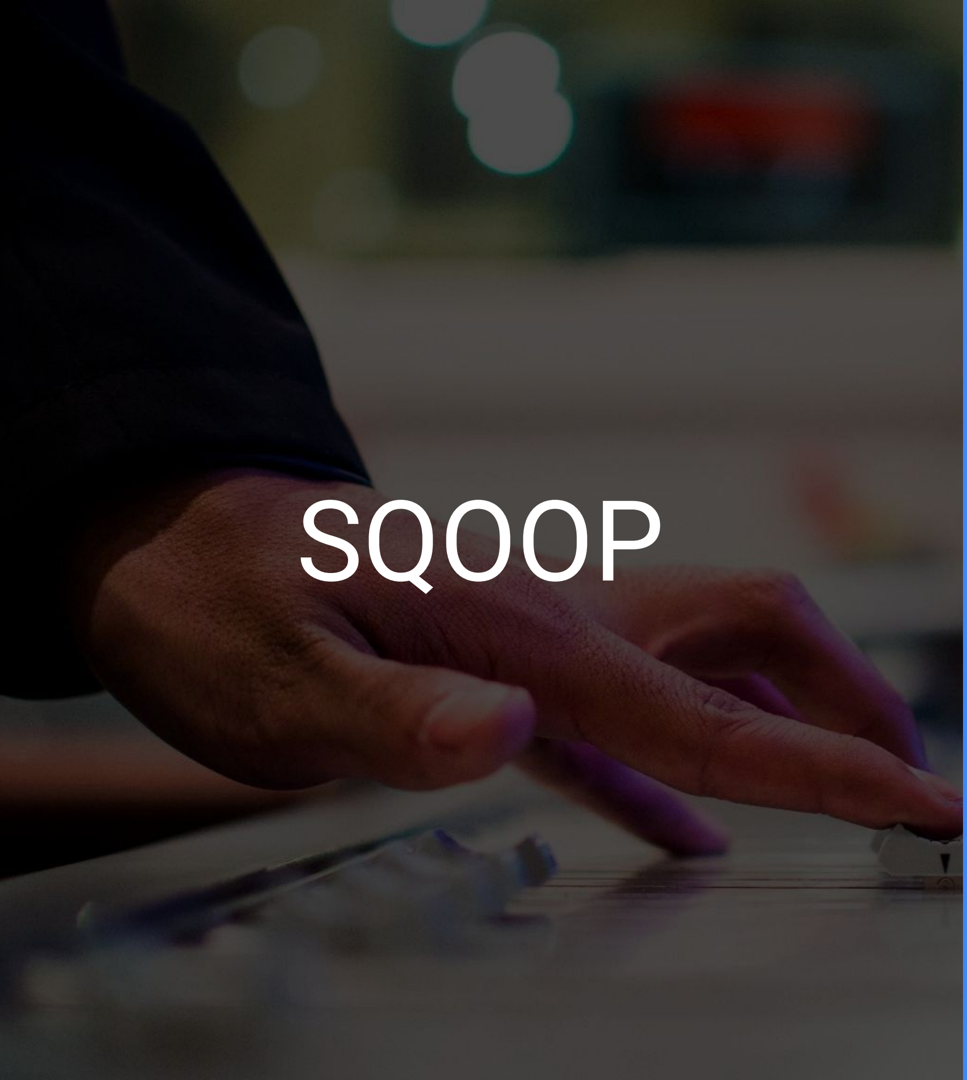


HBASE

Apache HBase is a Hadoop ecosystem component which is a distributed database that was designed to store structured data in tables that could have billions of row and millions of columns.

HBase is scalable, distributed, and NoSQL database that is built on top of HDFS.

HBase, provide real-time access to read or write data in HDFS.

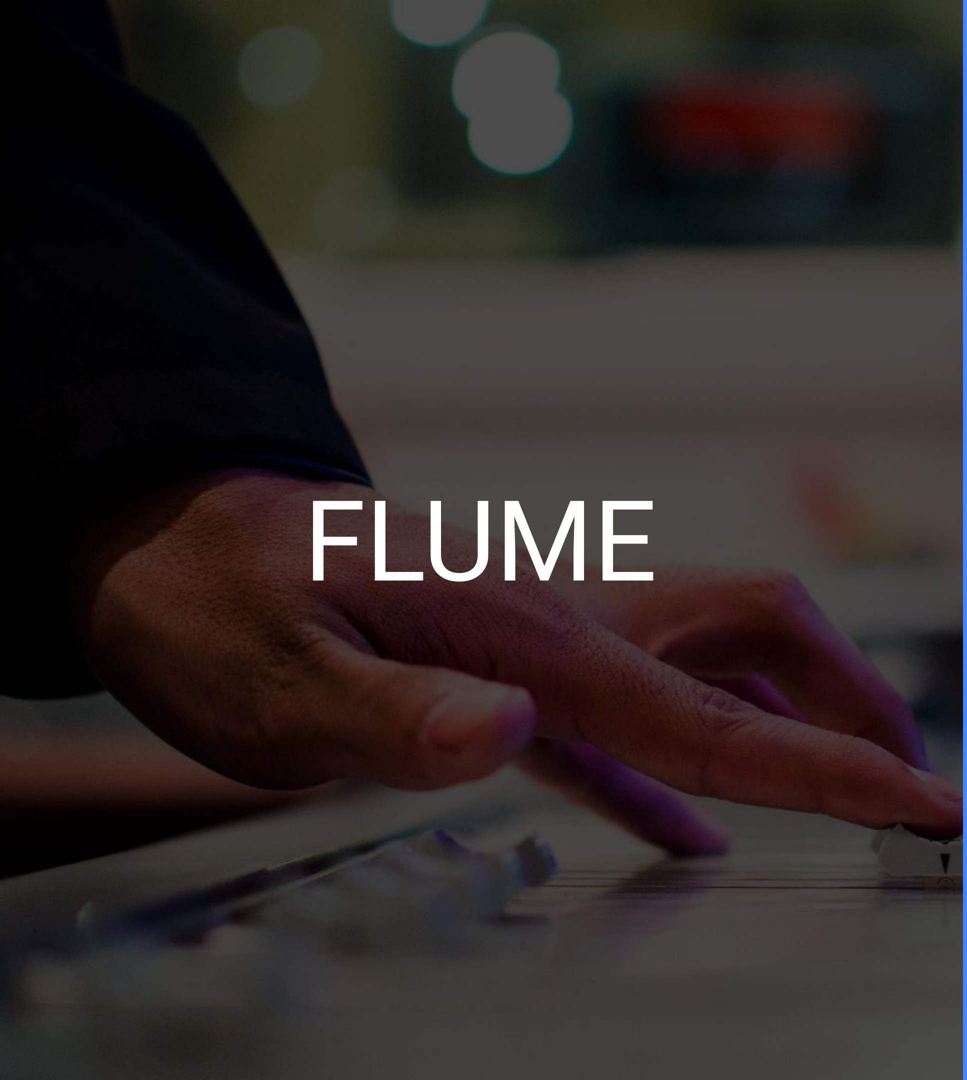


SQOOP

Sqoop imports data from external sources into related Hadoop ecosystem components like HDFS, Hbase or Hive.

It also exports data from Hadoop to other external sources.

Sqoop works with relational databases such as teradata, Netezza, oracle, MySQL



FLUME

Flume efficiently collects, aggregate and moves a large amount of data from its origin and sending it back to HDFS.

It is fault tolerant and reliable mechanism. This Hadoop Ecosystem component allows the data flow from the source into Hadoop environment.

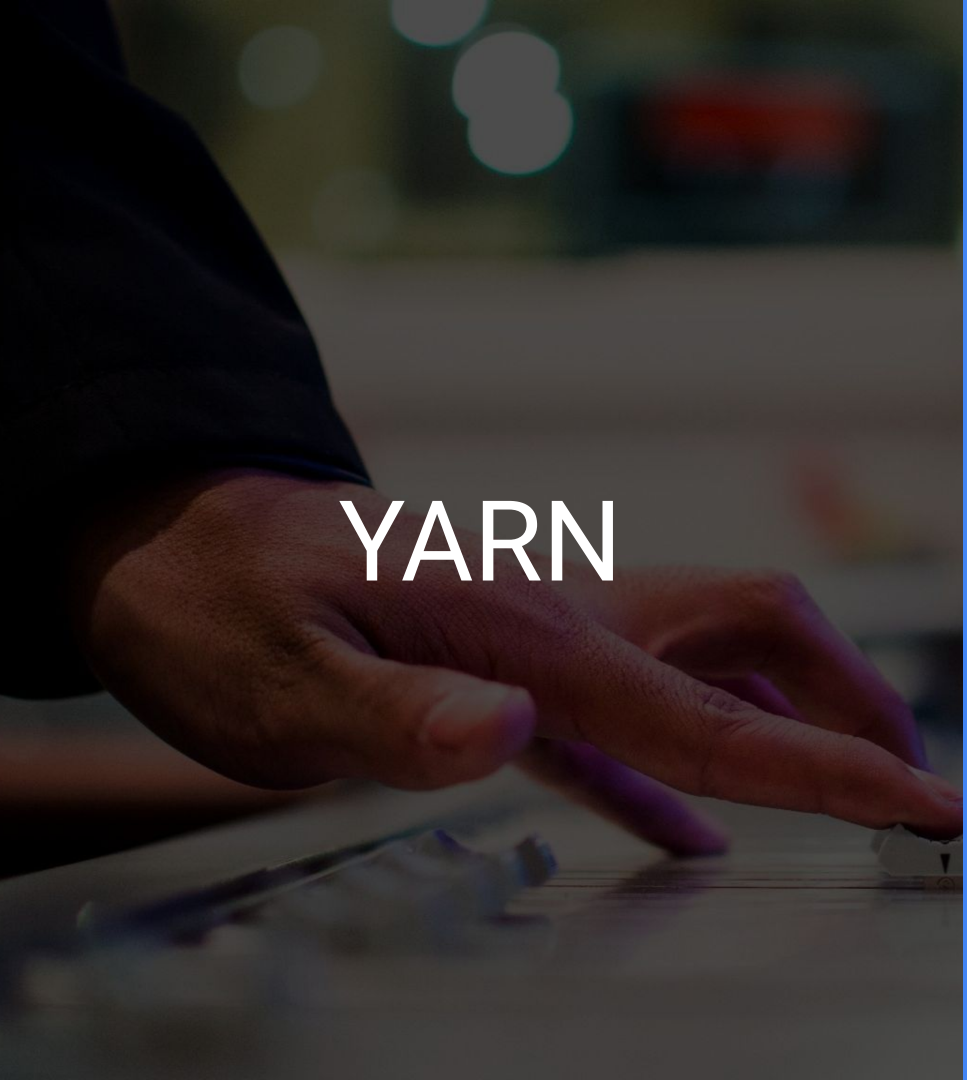
It uses a simple extensible data model that allows for the online analytic application. Using Flume, we can get the data from multiple servers immediately into hadoop.

ZOOKEEPER

Apache Zookeeper is a centralized service and a Hadoop Ecosystem component for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

Zookeeper manages and coordinates a large cluster of machines.

We are NOT going to learn the following Hadoop components.

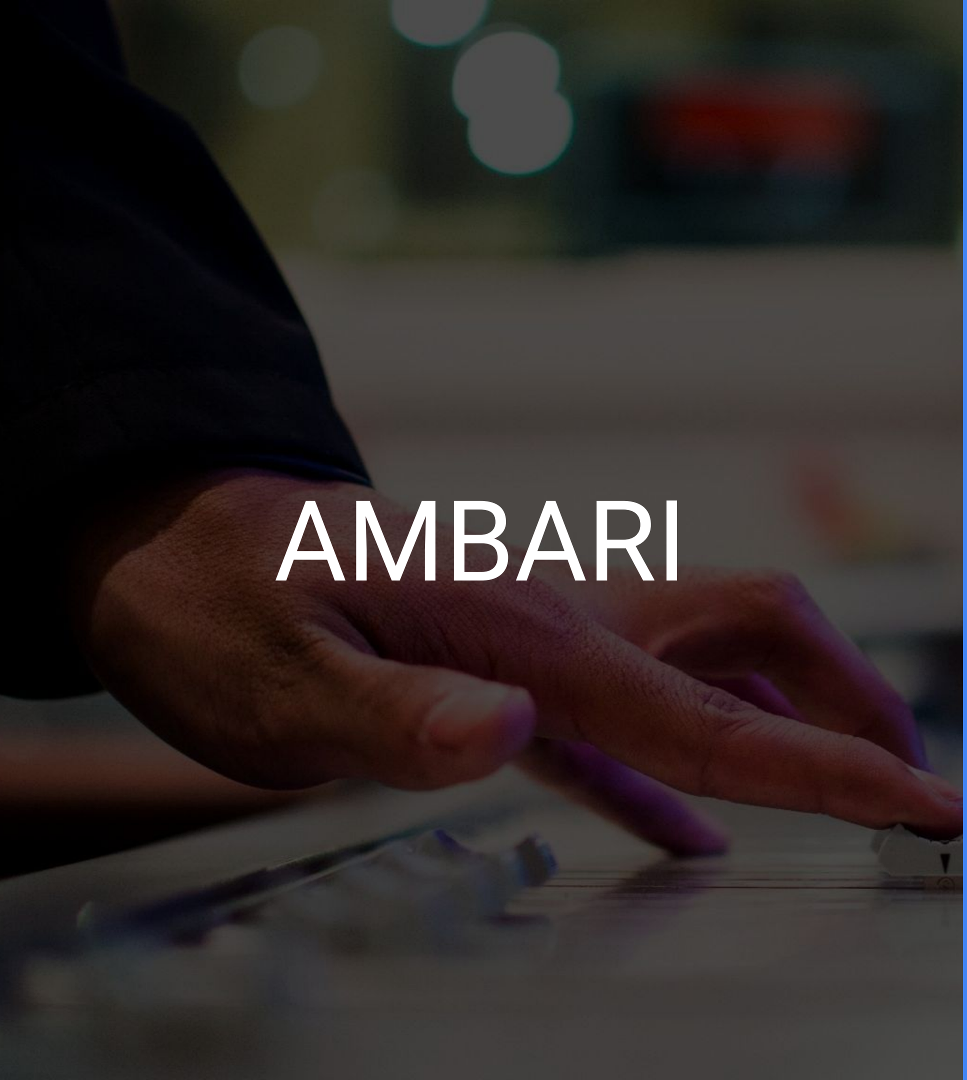


YARN

Hadoop YARN (Yet Another Resource Negotiator) is a Hadoop ecosystem component that provides the resource management.

YARN is also one the most important component of Hadoop Ecosystem.

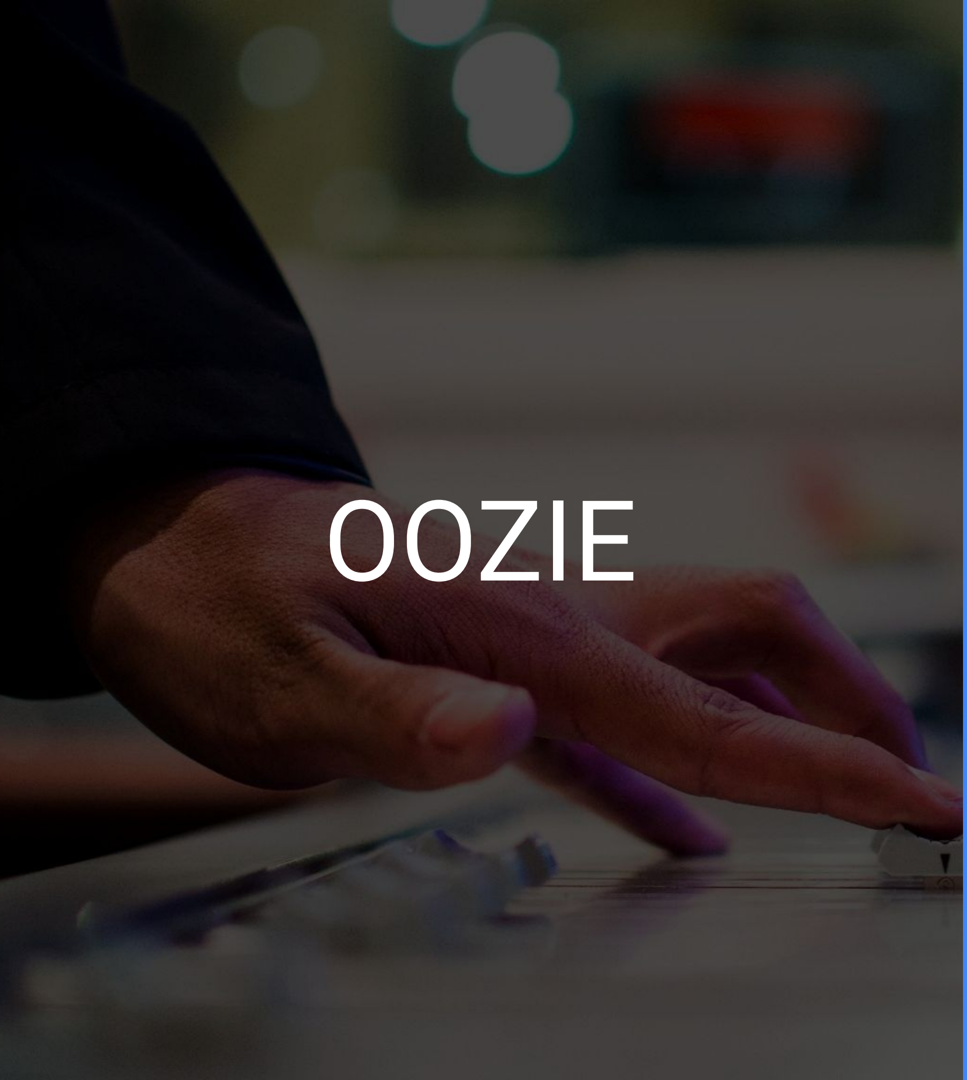
YARN is called as the operating system of Hadoop as it is responsible for managing and monitoring workloads. It allows multiple data processing engines such as real-time streaming and batch processing to handle data stored on a single platform.



AMBARI

Ambari, another Hadoop Ecosystem component, it’s a management platform for provisioning, managing, monitoring and securing apache Hadoop cluster.

Hadoop management gets simpler as Ambari provide consistent, secure platform for operational control.



OOZIE

Oozie is a workflow scheduler system for managing apache Hadoop jobs.

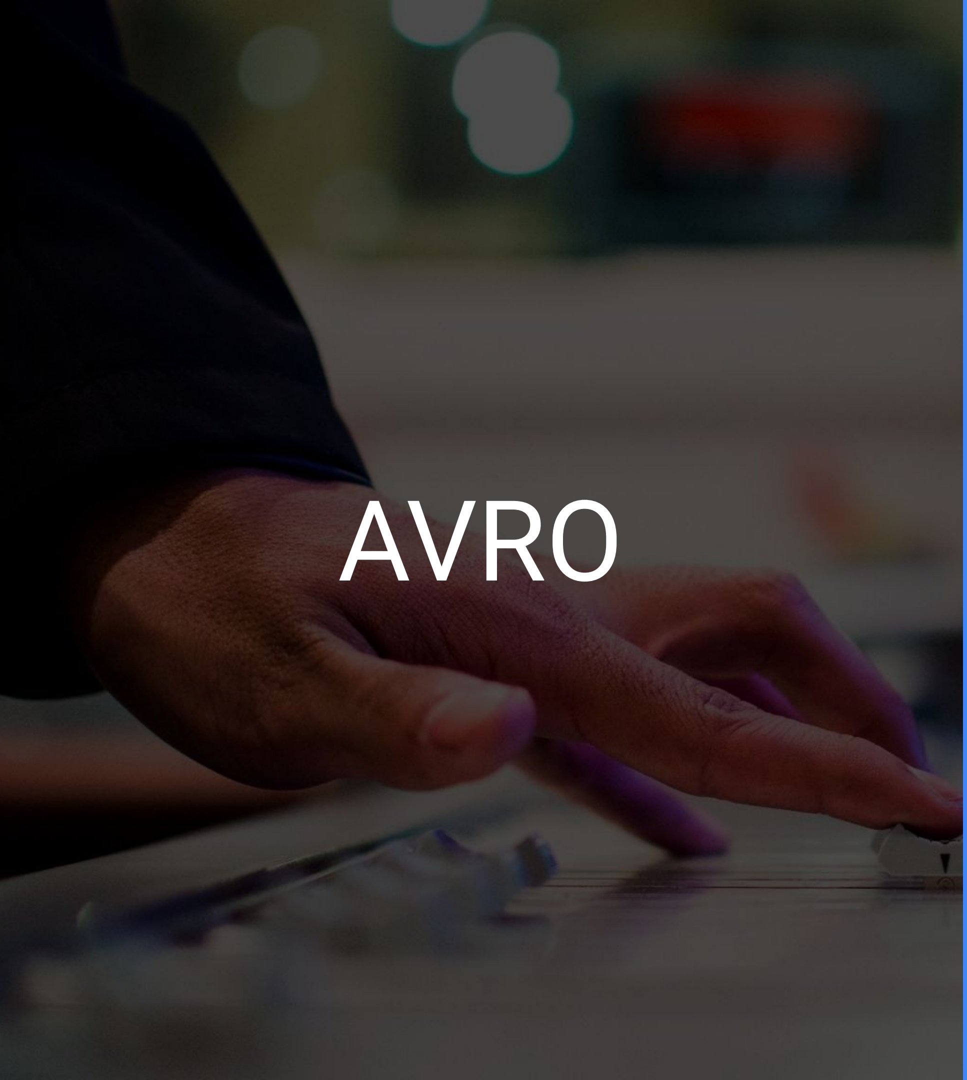
Oozie combines multiple jobs sequentially into one logical unit of work.

Oozie framework is fully integrated with apache Hadoop stack, YARN as an architecture center and supports Hadoop jobs for apache MapReduce, Pig, Hive, and Sqoop.

HCATALOG

It is a table and storage management layer for Hadoop. HCatalog supports different components available in Hadoop ecosystems like MapReduce, Hive, and Pig to easily read and write data from the cluster. HCatalog is a key component of Hive that enables the user to store their data in any format and structure.

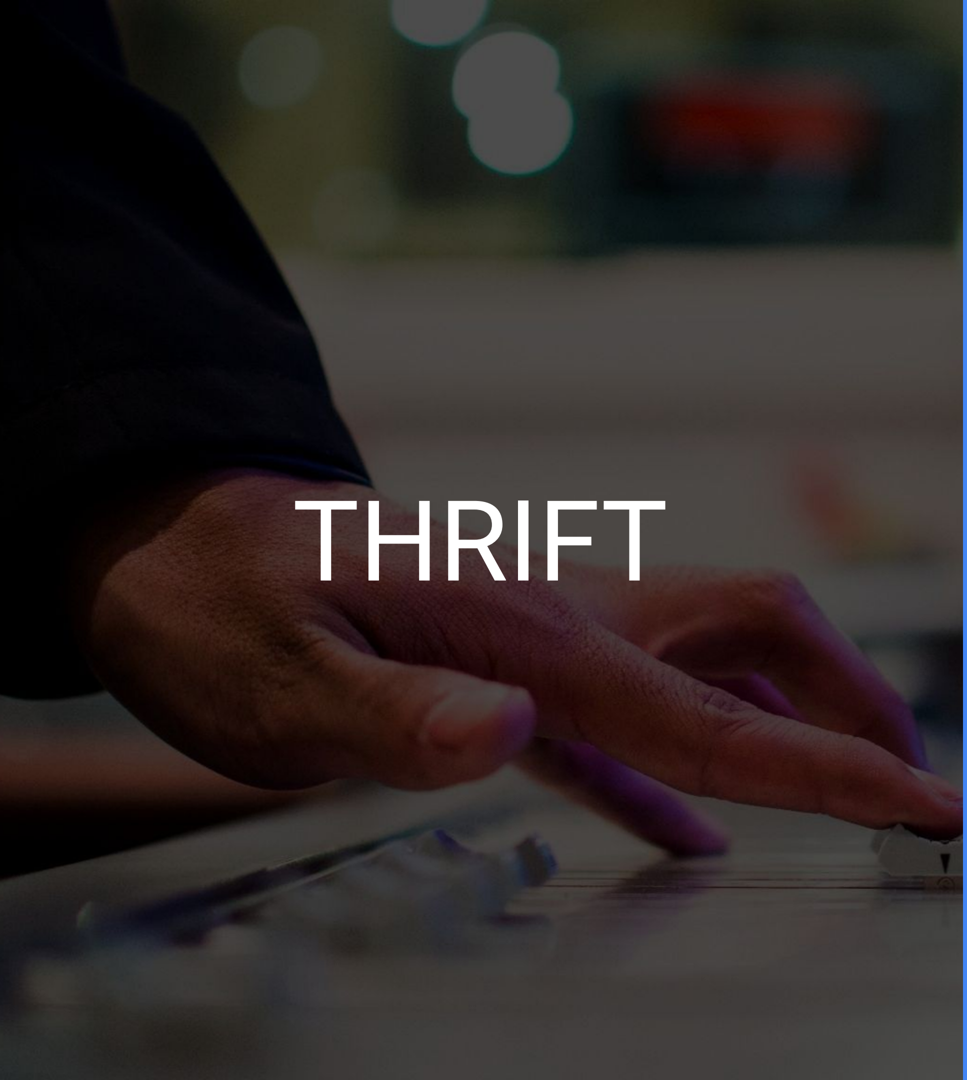
By default, HCatalog supports RCFile, CSV, JSON, sequenceFile and ORC file formats.



AVRO

Avro is a part of Hadoop ecosystem and is a most popular Data serialization system. Avro is an open source project that provides data serialization and data exchange services for Hadoop. These services can be used together or independently. Big data can exchange programs written in different languages using Avro.

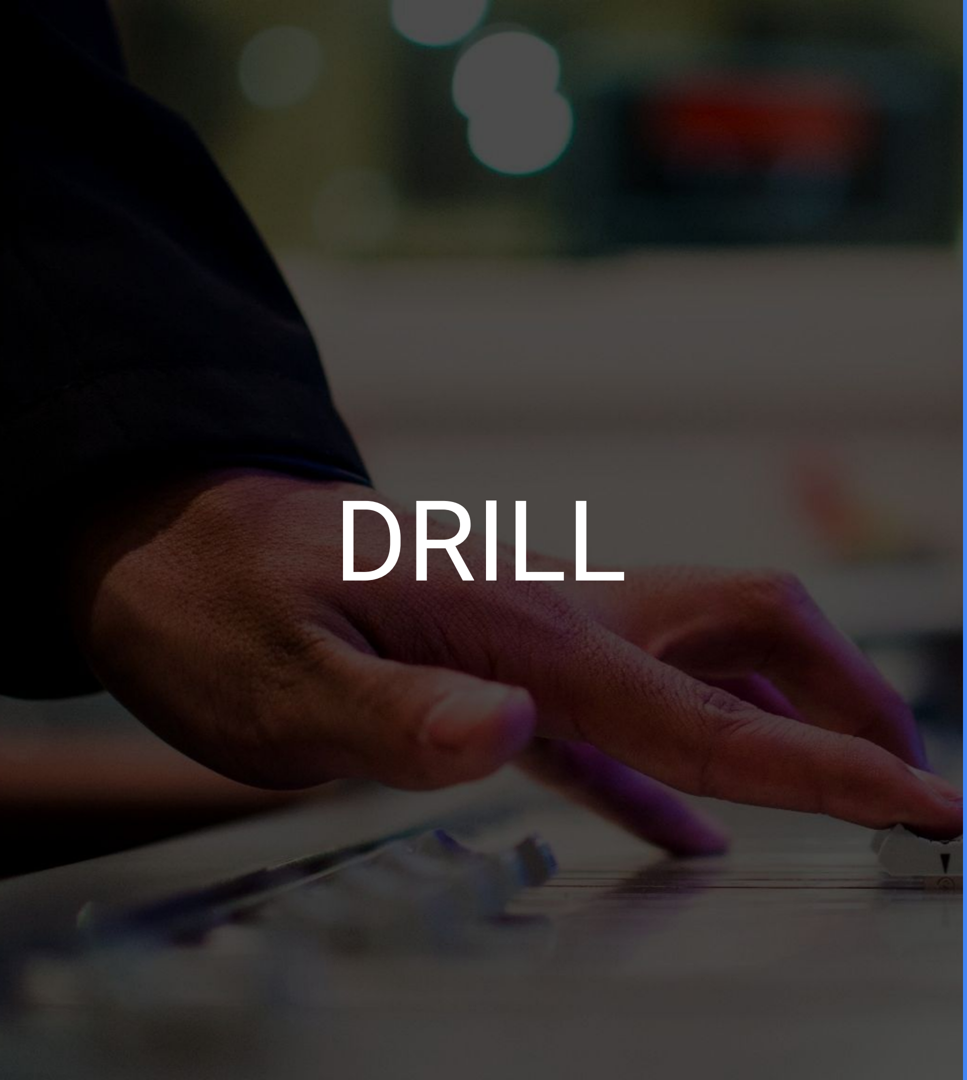
Using serialization service programs can serialize data into files or messages. It stores data definition and data together in one message or file making it easy for programs to dynamically understand information stored in Avro file or message.



THRIFT

It is a software framework for scalable cross-language services development. Thrift is an interface definition language for RPC(Remote procedure call) communication.

Hadoop does a lot of RPC calls so there is a possibility of using Hadoop Ecosystem componet Apache Thrift for performance or other reasons.



DRILL

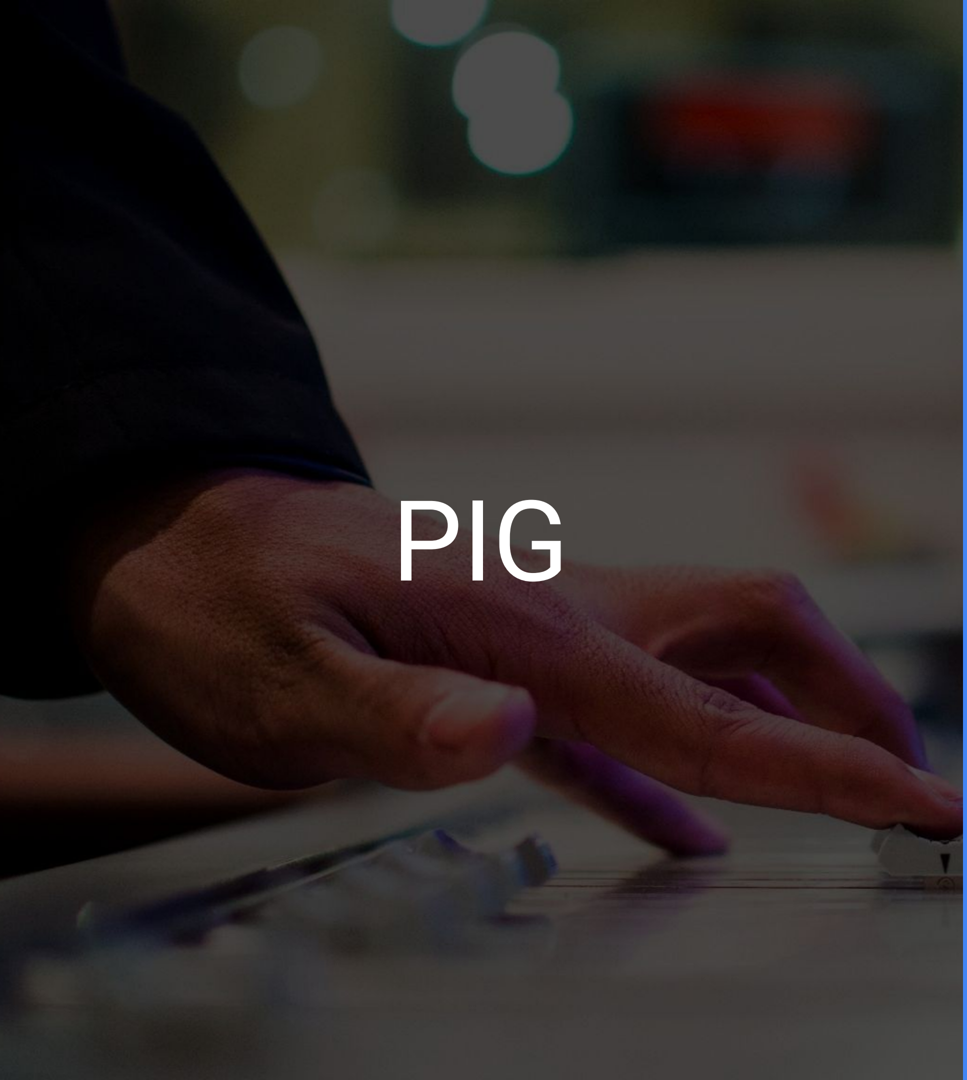
The drill is the first distributed SQL query engine that has a schema-free model.

The drill has become an invaluable tool at cardlytics, a company that provides consumer purchase data for mobile and internet banking. Cardlytics is using a drill to quickly process trillions of record and execute queries.

MAHOUT

Mahout is open source framework for creating scalable machine learning algorithm and data mining library.

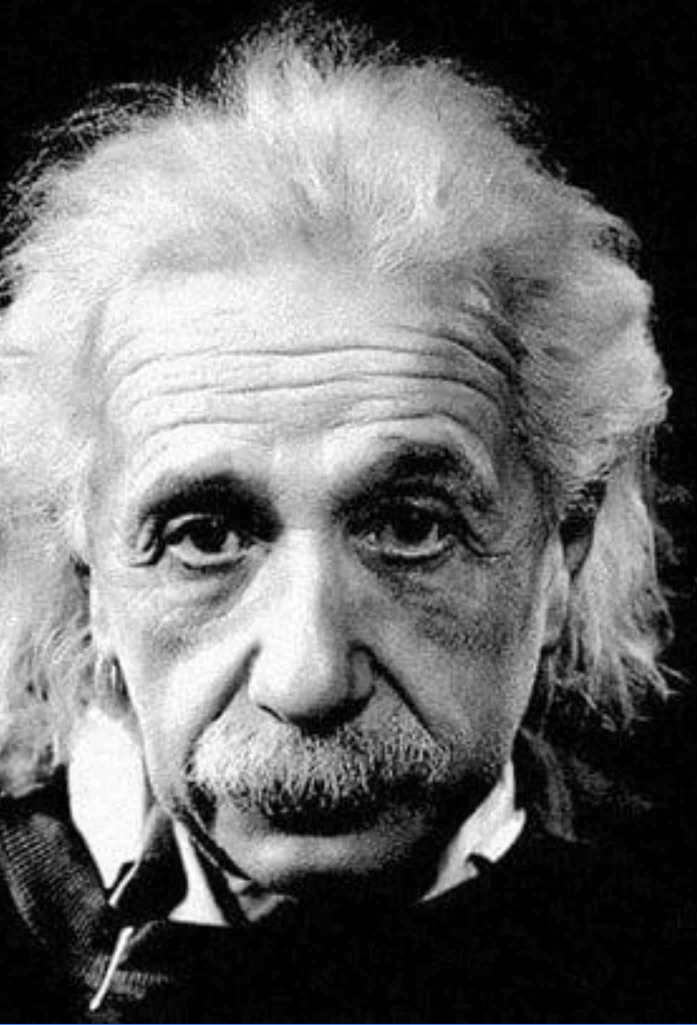
Once data is stored in Hadoop HDFS, mahout provides the data science tools to automatically find meaningful patterns in those big data sets.



PIG

Apache Pig is a high-level language platform for analyzing and querying huge dataset that are stored in HDFS. Pig as a component of Hadoop Ecosystem uses *PigLatin* language.

It is very similar to SQL. It loads the data, applies the required filters and dumps the data in the required format. For Programs execution, pig requires Java runtime environment.



“Intellectual growth should commence at birth and cease only at death.”

—Albert Einstein