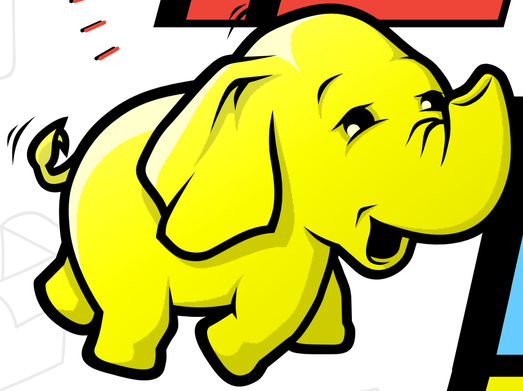
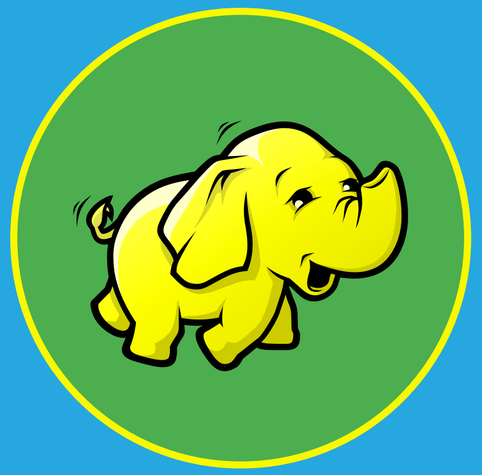
12 components of hadoop ECOSYSTEM



simplilearn



The popularity of Hadoop has grown in the last few years, because it meets the needs of many organizations for flexible data analysis capabilities with an unmatched price-performance curve.

The Hadoop ecosystem is continuously growing to meet the needs of Big Data. Let's understand the role of each component of the Hadoop ecosystem.

Hadoop Ecosystem comprises of the following 12 components:

Hadoop HDFS

SQOOP

Apache Spark

Pig

hadoop Hive

Oozie

HBase

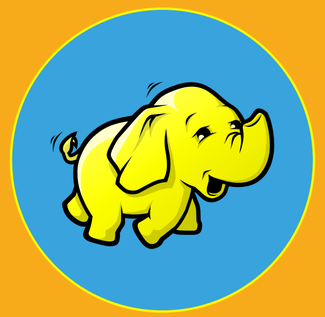
Flume

Hadoop MapReduce

Impala

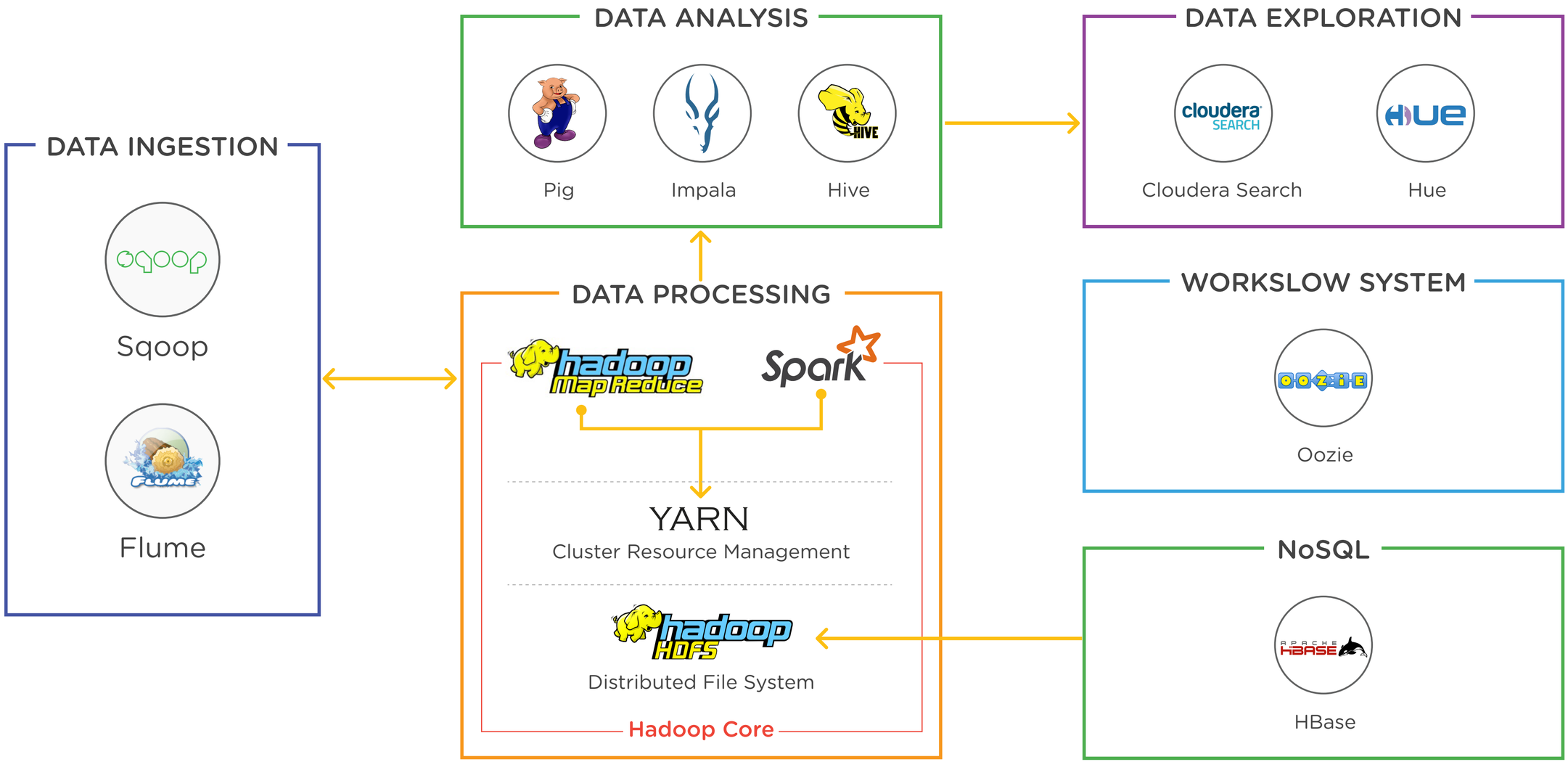
Cloudera Search

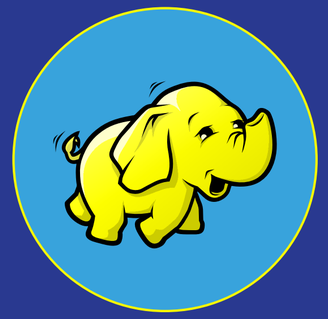
Hue



Now, let's check how these components work as part of an ecosystem

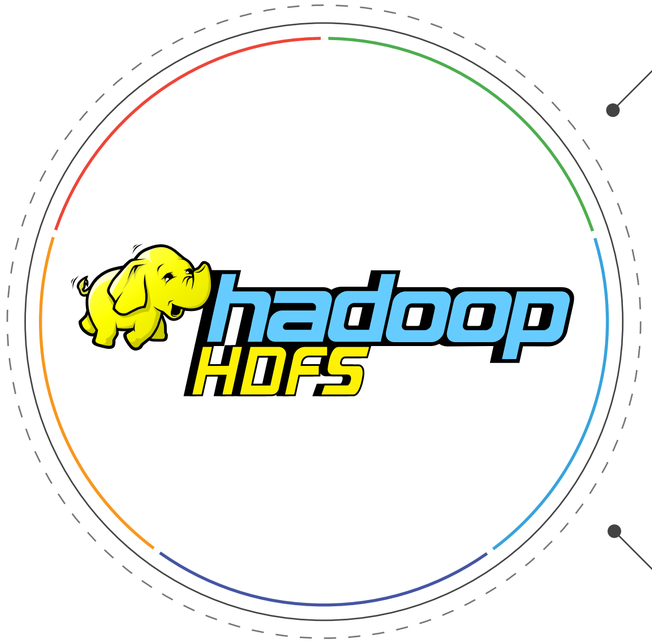
HADOOP ECOSYSTEM - COMPONENTS





What does each of these components actually do?

1 HADOOP DISTRIBUTED FILE SYSTEM



1. Provides file permissions and authentication
2. Streaming access to file system data
3. Hadoop provides a command line interface to interact with HDFS
4. Suitable for the distributed storage and processing
5. A storage layer for Hadoop

simplilearn

2 HBASE

Stores data in HDFS

A NoSQL database or non-relational database

The table can have thousands of columns

1 2 3 4 5

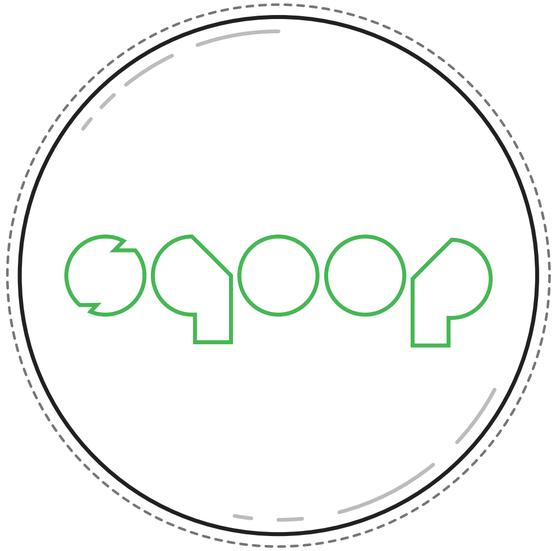
Mainly used when you need random, real-time, read/write access to your Big Data

Provides support to high volume of data and high throughput

simplilearn

3 SQOOP

Sqoop is a tool designed to transfer data between Hadoop and relational database servers



It is used to import data from relational databases such as Oracle and MySQL to HDFS and export data from HDFS to relational databases

simplilearn

4 FLUME

A distributed service for ingesting streaming data

Ideally suited for event data from multiple systems

simplilearn

5 SPARK

1. An open-source cluster computing framework
2. Supports Machine learning, Business intelligence, Streaming, and Batch processing
3. Provides 100 times faster performance than MapReduce

- Spark Core and Resilient Distributed Datasets (RDDs)
- Spark SQL
- Spark Streaming
- Machine Learning Library (Mlib)
- GraphX

simplilearn

6 HADOOP MAPREDUCE

Commonly used

1

2

An extensive and mature fault tolerance framework

Based on the map and reduce programming model

3

4

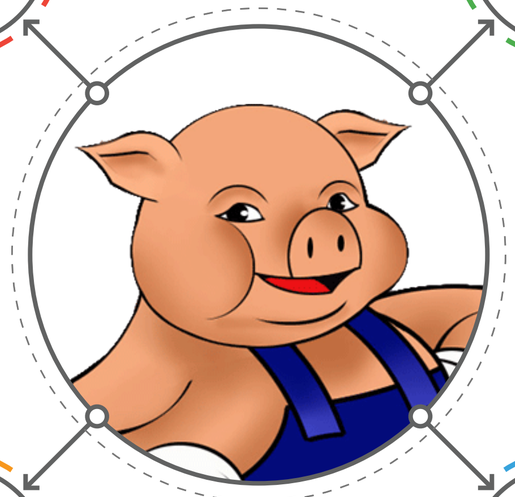
The original Hadoop processing engine which is primarily Java based

simplilearn

7 PIG

An open-source dataflow system

An alternative to writing Map-Reduce code



Converts pig script to Map-Reduce code

Best for ad-hoc queries such as join and filter

simplilearn

8 IMPALA

1

High performance SQL engine which runs on Hadoop cluster

2

Ideal for interactive analysis

3

Very low latency – measured in milliseconds

4

Supports a dialect of SQL (Impala SQL)

simplilearn

9 HIVE

1. Executes queries using MapReduce
2. Similar to Impala
3. Best for data processing and ETL

simplilearn

10 CLOUDERA SEARCH

One of Cloudera's near-real-time access products

Enables non-technical users to search and explore data stored in or ingested into Hadoop and HBase

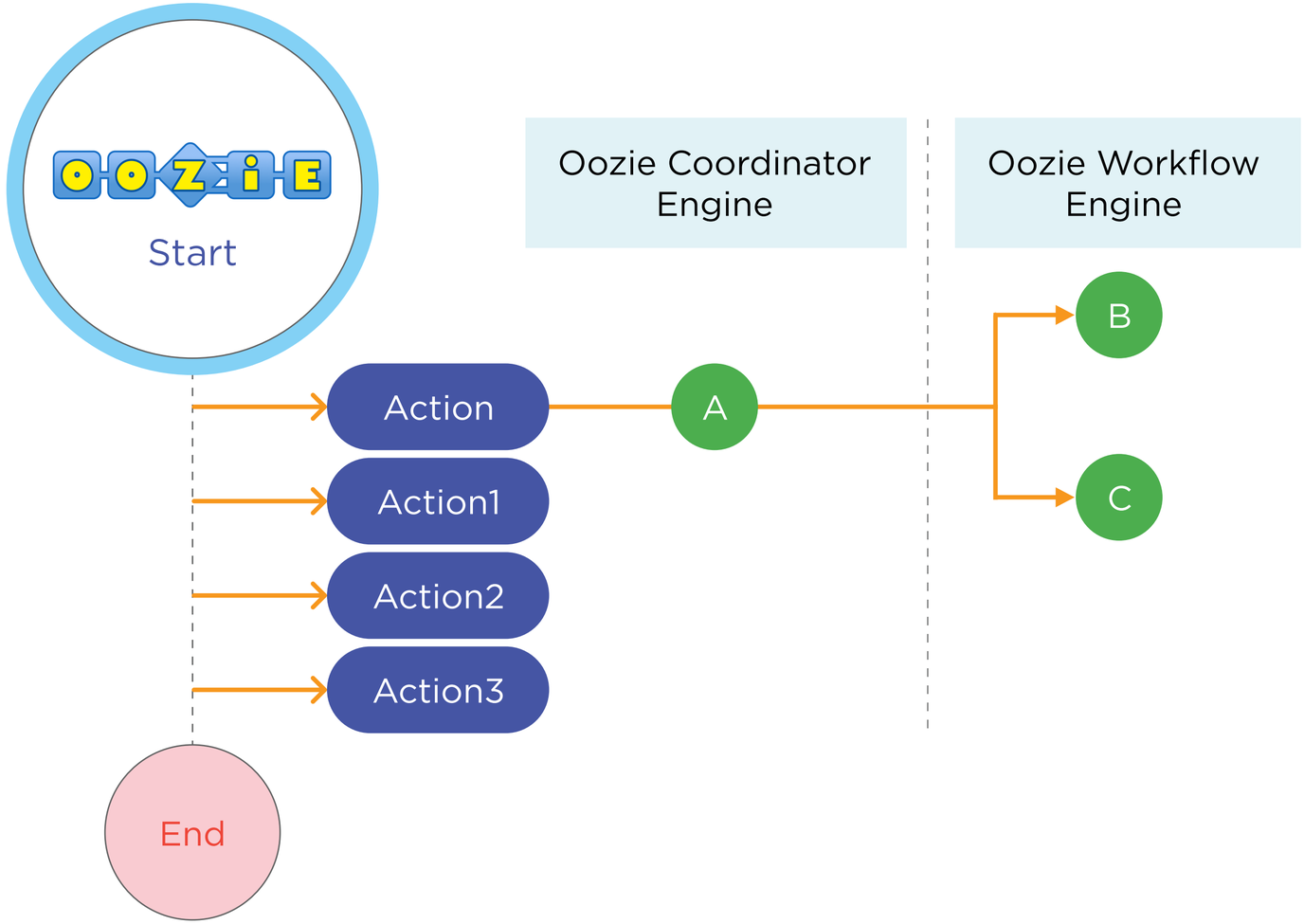
Users do not need SQL or programming skills to use Cloudera Search

A fully integrated data processing platform

simplilearn

11 OOZIE

Oozie is a workflow or coordination system used to manage the Hadoop jobs



simplilearn

12 HUE

It provides SQL editors for Hive, Impala, MySQL, Oracle, PostgreSQL, Spark SQL, and Solr SQL

Hue is an open source Web interface for analyzing data with Hadoop

simplilearn

DO YOU WANT TO BE

HADOOP CERTIFIED?

CLICK HERE

simplilearn

THANK YOU!

If you like this presentation,
please share it.

simplilearn