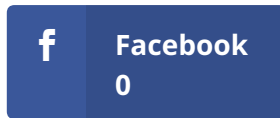




Hadoop Cheat Sheet

Jesse+ by | Mar 24, 2016 | Blog, Business, Data Engineering | 0 comments



Hadoop has a vast and vibrant developer community. Following the lead of Hadoop's name, the projects in the Hadoop ecosystem all have names that don't correlate to their function. This makes it really hard to figure out what each piece does or is used for. This is a cheat sheet to help you keep track of things. It is broken up into their respective general functions.

Distributed Systems

| Name | What It Is | What It Does | How It Helps |
|---------------------------|-------------------------------------|---|---|
| HDFS | A Distributed Filesystem for Hadoop | Acts as the filesystem or storage for Hadoop. | Improves the data input performance of MapReduce jobs with data locality. Creates a replicated, scalable file system. |
| Cassandra | A NoSQL database | A highly scalable database. | Allows you to scale a database in linear fashion. Can handle huge databases without bogging down. |
| | | | |

| | | | |
|-----------|---------------------------------------|---|--|
| HBase | A NoSQL database | Uses HDFS to create a highly scalable database. | Allows high scalability. Allows you to do random reads and writes with HDFS. |
| Zookeeper | A Distributed Synchronization Service | Provides synchronization of data amongst distributed nodes. | Allows a cluster to maintain consistent, distributed data across all nodes in a cluster. |

Processing Data



| Name | What It Is | What It Does | How It Helps |
|-----------|--|--|---|
| MapReduce | Distributed Programming Model and Software Framework | Breaks up a job into multiple tasks and processes them simultaneously. | Framework abstracts the difficult pieces of distributed systems. Allows vast quantities of data to be processed simultaneously. |
| Spark | General Purpose Processing Framework | Breaks up a job into multiple tasks and processes them simultaneously. | Framework abstracts the difficult pieces of distributed systems. Has more built-in functionality than MapReduce, like SQL. |
| Hive | Data Warehouse System | Allows use of query language to process data. | Helps SQL programmers harness MapReduce by creating SQL-like queries. |
| Pig | Data Analysis Platform | Processes data using a scripting language | Helps programmers use a scripting language to harness MapReduce power. |
| | | | |



| | | | |
|--------|-----------------------------------|---|--|
| Mahout | Machine Learning Library | Use a prewritten library to run machine learning algorithms on MapReduce. | Prevents you from having to rewrite machine learning algorithms to use MapReduce. Speeds up development time by using existing code. |
| Giraph | Graph Processing Library | Use a prewritten library to run graph algorithms on MapReduce. | Prevents you from having to rewrite graph algorithms to use MapReduce. Speeds up development time by using existing code. |
| MRUnit | Unit Test Framework for MapReduce | Run tests to verify your MapReduce job functions correctly. | Run programmatic tests to verify that a MapReduce program acts correctly. Has objects that allow you to mock up inputs and assertions to verify the results. |

Getting Data In/Out

| Name | What It Is | What It Does | How It Helps |
|-------|---------------------------|---|---|
| Avro | Data Serialization System | Gives an easy method to input and output data from MapReduce or Spark jobs. | Creates domain objects to store data. Makes easier data serialization and deserialization for MapReduce jobs. |
| Sqoop | Bulk Data Transfer | Moves data between Relational Databases and Hadoop. | Allows data dumps from the Relational Database to be placed |



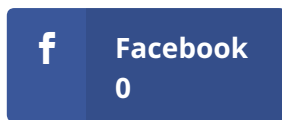
| | | | |
|-------|-------------------------------|---|---|
| | | | in Hadoop for processing later. Moves data output from a MapReduce job to be placed back in a Relational Database. |
| Flume | Data Aggregator | Handles large amounts of log data in a scalable fashion. | Moves large amounts of log data into HDFS. Since Flume scales so well, it can handle a lot of incoming data. |
| Kafka | Distributed Publish/Subscribe | Handles very high throughput and low latency message passing in a scalable fashion. | Decouples systems to allow many subscribers of published data. |

Administration

| Name | What It Is | What It Does | How It Helps |
|------------------|------------------------------------|--|--|
| Hue | Browser Based Interface for Hadoop | Allows users to interact with the Hadoop cluster over a web browser. | Makes it easier for users to interact with the Hadoop cluster. Granular permissions allow administrators to configure users' |
| Oozie | Workflow Engine for Hadoop | Makes creating complex workflows in Hadoop easier to create. | Allows you to create a complex workflow that leverages other projects like Hive, Pig, and MapReduce. Built-in logic allows users to handle failures of steps gracefully. |
| Cloudera Manager | Browser Based Manager for Hadoop | Allows easy configuration and configuration of | Eases the burden of dealing with and |

| | | | |
|--|--|-----------------|---|
| | | Hadoop cluster. | monitoring a large Hadoop Cluster. Helps install and configuration the Hadoop software. |
|--|--|-----------------|---|

The easiest way to install all of these programs is via [CDH](#) or Cloudera Distribution for Hadoop. The free edition of [Cloudera Manager](#) makes it even easier to create your cluster.



Become a Data Engineer

Are you tired of materials that don't go beyond the basics of data engineering?

Become a Data Engineer

AWS Screencast



GitHub