**Hadoop**

# Topics

* Capacity Scheduler
* Fair Scheduler
* Hive
* SQOOP
* Flume
* Oozie
* PIG
* HBase

Capacity Scheduler

The Capacity Scheduler was designed to allow organizations to share Hadoop clusters in a predictable and simple manner – primarily using the very common notion of ‘job queues’. It provides capacity guarantees for queues while providing elasticity for queues’ cluster utilization in the sense that unused capacity of a queue can be harnessed by overloaded queues that have a lot of temporal demand. This results in significantly higher cluster utilization while still providing predictability for Hadoop workloads.

For example, if you set up 5 queues, each queue would then have 20% of the total capacity for processing jobs. You can define the queue for which each job is assigned. For providing the necessary elasticity, the Capacity Scheduler allocates free resources to any queue beyond its guaranteed capacity. These excess resources can be reclaimed as necessary and assigned to other queues in order to meet capacity guarantees.

Fair Scheduler

FairScheduler, a pluggable scheduler for Hadoop that allows YARN applications to share resources in large clusters fairly. Fair scheduling is a method of assigning resources to applications such that all apps get, on average, an equal share of resources over time. Hadoop NextGen is capable of scheduling multiple resource types. By default, the Fair Scheduler bases scheduling fairness decisions only on memory. It can be configured to schedule with both memory and CPU.

Hive

Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. Apache Hive supports analysis of large datasets stored in Hadoop's HDFS and compatible file systems. This uses HiveQL based on SQL as a query language.

SQOOP

Sqoop is a command-line interface application for transferring data between relational databases and Hadoop It supports incremental loads of a single table or a free form SQL query as well as saved jobs which can be run multiple times to import updates made to a database since the last import. Imports can also be used to populate tables in Hive or HBase. Exports can be used to put data from Hadoop into a relational database.

Flume

Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS). It has a simple and flexible architecture based on streaming data flows; and is robust and fault tolerant with tunable reliability mechanisms for failover and recovery.

Oozie

Oozie is a workflow scheduler system to manage [Hadoop](http://en.wikipedia.org/wiki/Apache_Hadoop) jobs. It is a server-based Workflow Engine specialized in running workflow jobs with actions that run Hadoop MapReduce and Pig jobs. For the purposes of Oozie, a workflow is a collection of actions (e.g. Hadoop Map/Reduce jobs, Pig jobs) arranged in a control dependency. Oozie workflows contain control flow nodes and action nodes. Control flow nodes define the beginning and the end of a workflow (start, end and fail nodes) and provide a mechanism to control the workflow execution path (decision, fork and join nodes). Action nodes are the mechanism by which a workflow triggers the execution of a computation/processing task. Oozie provides support for different types of actions: Hadoop MapReduce, Hadoop file system, Pig, SSH, HTTP, eMail and Oozie sub-workflow.

PIG

Pig is a high-level platform for creating [MapReduce](http://en.wikipedia.org/wiki/MapReduce) programs used with [Hadoop](http://en.wikipedia.org/wiki/Hadoop). The language for this platform is called Pig Latin. Pig Latin abstracts the programming from the [Java](http://en.wikipedia.org/wiki/Java_(programming_language)) MapReduce idiom into a notation which makes MapReduce programming high level.

HBase

HBase is a non-relational and columnar database that allows for low-latency, quick lookups in Hadoop. It adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts and deletes. EBay and Facebook use HBase heavily.