Hadoop Ecosystem- Deployment And Management

Abhishek Tiwari

Citation: *A. Tiwari*, ”Hadoop Ecosystem- Deployment And Management”, Abhishek Tiwari, 2012. doi:10.59350/pxv5g-jg743

Published on: October 12, 2012

My notes and thoughts on Hadoop Ecosystem from book Hadoop Operations[1].

One of the major key take aways is emergence of the Hadoop cluster deployment and management tools such as hstack and Apache AMBARI. In our own setup we managed to deploy and scale the Hadoop clusters on AWS with few boto scripts and set of puppet recipies.

Apache Hadoop

Apache Hadoop is made up of two components

- A distributed filesystem called Hadoop Distributed Filesystem(HDFS) inspired from Google Filesystem(GFS).
- A computation layer called MapReduce that performs processing in parallel.

Apache Hive

- Allows developers to write a dialect of SQL, which in turn executed as one or more MapReduce jobs.
- Hive's dialect of SQL is called HiveQL which implements only a subset of SQL features.
- Hive works by defining a table like schema over an existing set of files in HDFS.
- Hive handles the glory details of extracting records from HDFS files when HiveQL query is performed.
- HiveQL queries are mapped to equivalent MapReduce jobs.
- Using user-defined functions developers can extend the Hive functionality.

Apache Pig

- Like Apache Hive, Pig simplifies the authoring of MapReduce jobs.
- Developers write data processing jobs in a high-level scripting language.
- Pig converts these script in an execution plan and execute a series of MapReduce jobs.
- Using user-defined functions developers can extend Pigs built-in operations.

Apache Sqoop

- Also know as "SQL to Hadoop"
- Performs bi-directional data transfer between Hadoop and almost any RDBMS.
- Using MapReduce, Sqoop performs these operations in parallel.

[1] Hadoop Operations

- To connect RDBMS Sqoop uses JDBC driver or database specific plugins.
- Database specific plug-ins use native features of RDBMS and have great performance than JDBC drivers.
- Current native connectors include MySQL and PostgreSQL

Apache Flume

- Distributed streaming data collection and aggregation system.
- Moves massive volumes of data into systems such as Hadoop.
- Supports native connectivity and support for writing directly to HDFS.
- Streaming data delivery from a variety of sources including RPC, log4j, syslog.
- Data can be routed, load-balanced, replicated to multiple destinations and aggregated from thousands of hosts by a tier of agents.

Apache Oozie

- A workflow engine and scheduler built specifically for large-scale job orchestration on a Hadoop cluster.
- Allows to run many coordinated MapReduce jobs in workflow.
- Developers can use REST service for programmatic management of workflows and status retrieval.

Apache Whirr

- Provide a set of libraries to create and deploy Hadoop clusters in cloud environment.
- Uses jclouds library
- Apart from Hadoop, can provision Cassandra, HBase etc as service.

Apache HBase

- A low-latency, distributed, non-relational database built on top of HDFS.
- Modeled after Google's Bigtable.
- A flexible data model with scale-out properties and a very simple API.
- Data is stored in a semi-columnar format partitioned by rows into regions.

Apache Ambari[2]

- Web-based tool for installing, managing and monitoring Hadoop clusters.
- Provides an easy-to-use, step-by-step wizard for installing Hadoop services.
- Supports HBase, Hadoop, Hive, Oozie, Pig, Sqoop, Zookeeper.
- Leverages Puppet to perform installation and configuration of Hadoop services for the cluster.
- Provides central management for starting, stopping, and reconfiguring Hadoop services.
- Provides monitoring health and status of Hadoop cluster.
- Leverages Ganglia to collect system metrics and Nagios to monitor and trigger alerts.
- Supports RHEL/CentOS, Ubuntu
- Works with AWS IP Addresses

Hstack[3]

- Puppet recipes for deploying and configuring Hadoop/HBase clusters.
- Open sourced by Adobe.
- For more details see(https://github.com/hstack/puppet).

Apache Mesos[4]

- Cluster manager that allows to run Hadoop, MPI, Spark and other applications on dynamically shared pool of nodes.
- Run multiple instances/versions of Hadoop/application on the same cluster to isolate production from test
- Isolation between tasks with Linux containers

[2] Managing Hadoop Clusters with AMBARI
[3] Puppet recipes for deploying Hstack
[4] Apache Mesos

10.59350/pxv5g-jg743