PIG:

Java. Pig’s simple SQL-like scripting language is called Pig Latin, and appeals to developers already familiar with scripting languages and SQL.

Pig is complete, so you can do all required data manipulations in Apache Hadoop with Pig. Through the User Defined Functions(UDF) facility in Pig, Pig can invoke code in many languages like JRuby, Jython and Java. You can also embed Pig scripts in other languages. The result is that you can use Pig as a component to build larger and more complex applications that tackle real business problems.

Pig works with data from many sources, including structured and unstructured data, and store the results into the Hadoop Data File System.

Pig scripts are translated into a series of MapReduce jobs that are run on the Apache Hadoop cluster.

**SQOOP:**

Efficiently transfers bulk data between Apache Hadoop and structured datastores

Apache Sqoop efficiently transfers bulk data between Apache Hadoop and structured datastores such as relational databases. Sqoop helps offload certain tasks (such as ETL processing) from the EDW to Hadoop for efficient execution at a much lower cost. Sqoop can also be used to extract data from Hadoop and export it into external structured datastores. Sqoop works with relational databases such as Teradata, Netezza, Oracle, MySQL, Postgres, and HSQLDB

What Sqoop Does

Apache Sqoop does the following to integrate bulk data movement between Hadoop and structured datastores:

YARN coordinates data ingest from Apache Sqoop and other services that deliver data into the Enterprise Hadoop cluster.

**HIVE**

Hive provides a SQL-like interface to data stored in HDP. In the previous tutorial, we used Pig, which is a scripting language with a focus on dataflows. Hive provides a database query interface to Apache Hadoop.

## HIVE OR PIG?

People often ask why do Pig and Hive exist when they seem to do much of the same thing. Hive because of its SQL like query language is often used as the interface to an Apache Hadoop based data warehouse. Hive is considered friendlier and more familiar to users who are used to using SQL for querying data. Pig fits in through its data flow strengths where it takes on the tasks of bringing data into Apache Hadoop and working with it to get it into the form for querying. A good overview of how this works is in Alan Gates posting on the Yahoo Developer blog titled Pig and Hive at Yahoo!. From a technical point of view, both Pig and Hive are feature complete, so you can do tasks in either tool. However, you will find one tool or the other will be preferred by the different groups that have to use Apache Hadoop. The good part is they have a choice and both tools work together.

**HBASE**

A non-relational (NoSQL) database that runs on top of HDFS

Apache HBase is an open source NoSQL database that provides real-time read/write access to those large datasets.

HBase scales linearly to handle huge data sets with billions of rows and millions of columns, and it easily combines data sources that use a wide variety of different structures and schemas. HBase is natively integrated with Hadoop and works seamlessly alongside other data access engines through YARN.

### What HBase Does

Apache HBase provides random, real time access to your data in Hadoop. It was created for hosting very large tables, making it a great choice to store multi-structured or sparse data. Users can query HBase for a particular point in time, making “flashback” queries possible. These following characterisitcs make HBase a great choice for storing semi-structured data like log data and then providing that data very quickly to users or applications integrated with HBase.

Enterprises use Apache HBase’s low latency storage for scenarios that require real-time analysis and tabular data for end user applications. One company that provides web security services maintains a system accepting billions of event traces and activity logs from its customer’ desktops every day. The company’s programmers can tightly integrate their security solutions with HBase (to assure that the protection they provide keeps pace with real-time changes in the threat landscape.)

Another company provides stock market ticker plant data that its users query more than thirty thousand times per second, with an SLA of only a few milliseconds. Apache HBase provides that super low-latency access over an enormous, rapidly changing data store.

**Flume**

 A service for streaming logs into Hadoop

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.

YARN coordinates data ingest from Apache Flume and other services that deliver raw data into an Enterprise Hadoop cluster.

### What Flume Does

Flume lets Hadoop users ingest high-volume streaming data into HDFS for storage. Specifically, Flume allows users to:

Enterprises use Flume’s powerful streaming capabilities to land data from high-throughput streams in the [Hadoop Distributed File System (HDFS)](https://hortonworks.com/hadoop/hdfs/" \o "Hadoop Distributed File System (HDFS)). Typical sources of these streams are application logs, sensor and machine data, geo-location data and social media. These different types of data can be landed in Hadoop for future analysis using interactive queries in Apache Hive. Or they can feed business dashboards served ongoing data by Apache HBase.

In one specific example, Flume is used to log manufacturing operations. When one run of product comes off the line, it generates a log file about that run. Even if this occurs hundreds or thousands of times per day, the large volume log file data can stream through Flume into a tool for same-day analysis with Apache Storm or months or years of production runs can be stored in HDFS and analyzed by a quality assurance engineer using Apache Hive.