**OVERVIEW**

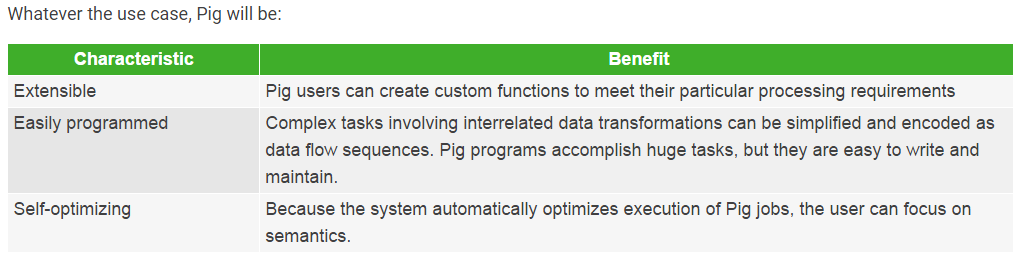
A scripting platform for processing and analyzing large data sets

With YARN as the architectural center of ApacheTM Hadoop, multiple data access engines such as Apache Pig interact with data stored in the cluster. Apache Pig allows Apache Hadoop users to write complex MapReduce transformations using a simple scripting language called Pig Latin. Pig translates the Pig Latin script into MapReduce so that it can be executed within YARN for access to a single dataset stored in the Hadoop Distributed File System (HDFS).

**WHAT PIG DOES**

Pig was designed for performing a long series of data operations, making it ideal for three categories of Big Data jobs:

* **Extract-transform-load (ETL)** data pipelines,
* **Research** on raw data, and
* **Iterative data processing**



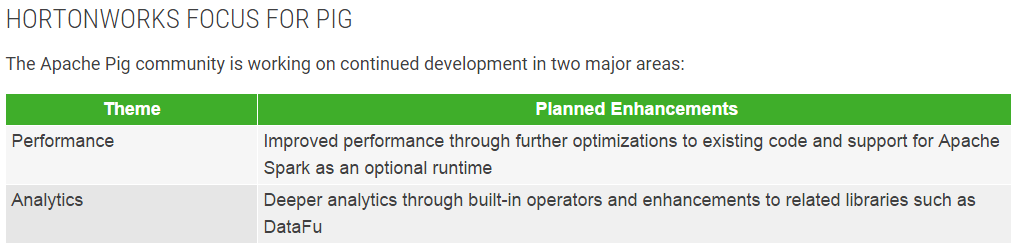
**HOW PIG WORKS**

Pig runs on Apache Hadoop YARN and makes use of MapReduce and the Hadoop Distributed File System (HDFS). The language for the platform is called Pig Latin, which abstracts from the Java MapReduce idiom into a form similar to SQL. While SQL is designed to query the data, Pig Latin allows you to write a data flow that describes how your data will be transformed (such as aggregate, join and sort).

Since Pig Latin scripts can be graphs (instead of requiring a single output) it is possible to build complex data flows involving multiple inputs, transforms, and outputs. Users can extend Pig Latin by writing their own functions, using Java, Python, Ruby, or other scripting languages. Pig Latin is sometimes extended using UDFs (User Defined Functions), which the user can write in any of those languages and then call directly from the Pig Latin.

The user can run Pig in two modes, using either the “pig” command or the “java” command:

* **MapReduce Mode.** This is the default mode, which requires access to a Hadoop cluster.
* **Local Mode.** With access to a single machine, all files are installed and run using a local host and file system.



**HDP 2.4 – Pig 0.15.0**

* Pig on Tez stablization
* Improved Tez auto-parallelism
* Ability to invoke Hive UDFs from Pig

