The Pig platform is a relatively easy tool for creating Apache MapReduce applications. With an active open-source community contributing to the project, Pig is rapidly gaining ground as a high-level data flow programming language and execution framework for analyzing big data.

The Pig platform consists of the Pig Latin programming language and the infrastructure to support applications that run on distributed file systems:

- The infrastructure layer includes a compiler that produces MapReduce programs.
- Pig Latin is a high-level textual language that makes it easy to write, understand, and maintain programs. Practically, this means you can write a Pig script in 15 minutes that might have taken you hours to write in the Java language. The Pig platform also optimizes task execution automatically and can be extended with custom functions.

The Pig system sits on top of the Apache Hadoop Distributed File System (Apache HDFS). It reads and writes data from HDFS as well as other locations and runs MapReduce jobs. Essentially, the Pig platform is a way to interface with the Hadoop framework easily.

Pig Latin: Why Another Programming Language?

Pig Latin is a nontraditional programming language that focuses on data flow rather than the traditional programming operations used by languages such as Java or Python. As a Pig Latin user, you build a script by specifying one or more input data sets, and then identifying the operations to apply. These operations might include filtering out unwanted records, joining two data sets, and splitting a data set based on certain criteria. You connect the specified operations in a graph that automatically finds the optimal data flow. Then you define how to write out the data to one or more output sources. This is a very different way to program.

Pig manages all the translation and coordination with the Hadoop framework for you. First, it automatically performs optimizations on the Pig Latin script, and then it translates the required operations into one or more—typically 10, 20, or even 50—MapReduce jobs. Pig runs those jobs in the Hadoop cluster, while at the same time providing feedback on progress, error messages, and so on.

---

“Pig Latin is ideal for users who don’t want to work with more complex Java code to create MapReduce applications.”

—Alan Gates
Pig versus Apache Hive

Users often look at the Pig platform and the Apache Hive data warehouse infrastructure as different tools to accomplish the same goals. However, Pig and Hive are actually going after different analytics sweet spots. Hive presents a Structured Query Language (SQL) layer on top of MapReduce, which works well for answering questions from structured relational data. The Pig language is built to work in the Hadoop world, where data may or may not be structured or have known schemas.

The Pig platform is built to provide a pipeline of operations. It is often used to do extract, transform, load (ETL) operations by pulling data from various sources and then doing the transformations in the Hadoop framework as defined in the user’s script. For example, using SQL or Hive, you might ask, “How many people who came to my web site yesterday also came back today?” In Pig you can use a single script to define a much longer set of operations, such as, “Find all the people who came to my web site yesterday and today, analyze what they bought, and figure out their demographic group and geographic region.”

Pig Limitations

While the Pig platform is designed for ETL-type use cases, it’s not a great choice for real-time scenarios. Pig is built on top of MapReduce, which is batch oriented. Pig is also not the right choice for pinpointing a single record in very large data sets. Pig works best processing the entire data set.

How Easy Is Pig Latin Really?

Pig Latin is ideal for users who don’t want to work with more complex Java code to create MapReduce applications. If you’ve worked with data operations before, you will have no trouble learning the Apache Pig language. There aren’t an overwhelming number of operations that you typically apply to data. You may need to join data, filter it, protect it, group it, sort it, and so on. But the Pig language enables all those actions at a high level, so you won’t have to learn a lot to start writing your Pig script. A new user can write a first script in 30 minutes and start using the program. However, the Pig platform is a highly complex tool, and you can do a lot more with it as you gain more experience.

How Does the Pig Platform Handle Relational Systems Data?

Apache Pig can work with relational data sets in two ways:

- You can load relational data directly into the Apache Hadoop framework, where Pig can access it.
- Using database connectors, Pig can load data directly from a relational database system and access it from there.

The default operation for the Pig platform is to look first to read and write data on the Apache Hadoop Distributed File System (HDFS), but because it can also read from the Apache HBase database and relational databases, it’s possible to pull data from multiple sources for processing. For example, the bulk of the data might be in HDFS, but Pig could also be reading some data out of a relational system and using that to process the data in Hadoop.

1 The Apache HBase database is a scalable, distributed database that supports structured data storage for large tables.
Next Steps for the Pig Platform

Pig is a very active project. Developers from at least five different companies are committed to contributing work. Some of the areas we’d like to see the project grow include:

- **Taking advantage of Hadoop YARN.** Hadoop YARN is a new framework for job scheduling and cluster resource management that enables developers to build tools other than MapReduce applications for the Hadoop framework. YARN will be part of the Hadoop 2.0 release and presents opportunities for growing the Pig platform.

- **Expanding integration with other scripting languages.** Rather than be constrained by the operators the Pig language offers, you can use a user-defined function to write your Pig Latin script in the Python and Ruby languages, as well as the Java language. Continuing to expand language sets will make Pig more flexible and enable users to write in the language of their choice.

- **Improve the user experience.** The Pig platform is built by engineers, and sometimes that shows. More work on making the user experience friendly will make the Pig platform a better tool.

This paper is derived from an interview with Alan Gates on August 28, 2012. For the full interview, [listen to the podcast](#).

For more information about the Apache Pig project, visit [http://pig.apache.org](http://pig.apache.org).