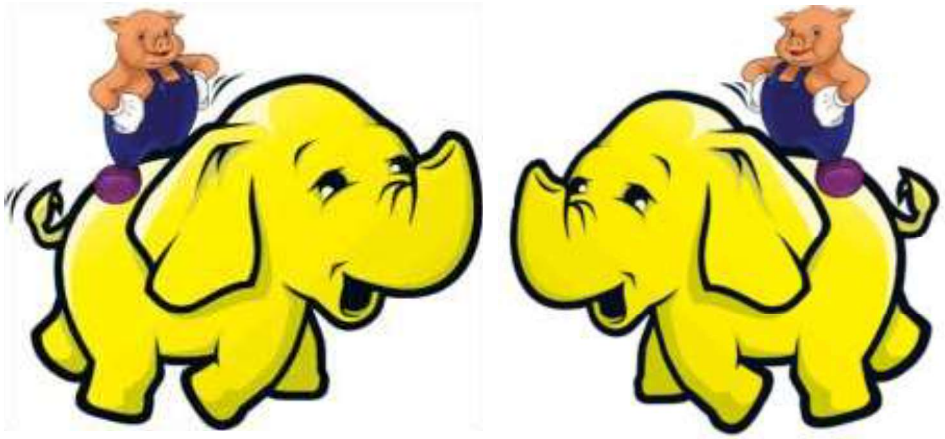
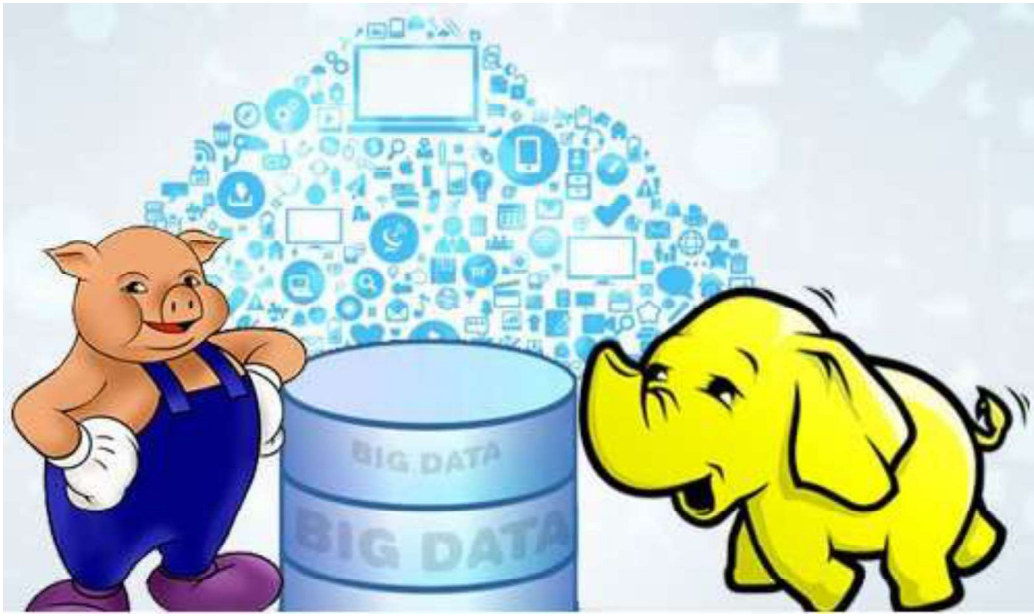
रा.इ.सू.प्रौ.सं
NIELIT



National Institute of Electronics and Information Technology

Big Data and Hadoop

Module 6-Power of Pig



History

1. In the summer of 2006, Pig started out as a research project in Yahoo! Research, where Yahoo! scientists designed it and produced an initial implementation. The researchers felt that the MapReduce paradigm presented by Hadoop "is too low-level and rigid, and leads to a great deal of custom user code that is hard to maintain and reuse."
2. At the same time they observed that many MapReduce users were not comfortable with declarative languages such as SQL. Thus they set out to produce "a new language called Pig Latin that we have designed to fit in a sweet spot between the declarative style of SQL, and the low-level, procedural style of MapReduce."
3. About this same time, in fall 2007, Pig was open sourced via the Apache Incubator. The first Pig release came a year later in September 2008. Later that same year, Pig graduated from the Incubator and became a subproject of Apache Hadoop.
4. Early in 2009 other companies started to use Pig for their data processing. By the end of 2009 about half of Hadoop jobs at Yahoo! were Pig jobs. Apart from Yahoo, Twitter, AOL, LinkedIn, Nokia, PayPal, Salesforce.com etc. are some of the major companies that are using Apache Pig.

Apache Pig Philosophy

What does it mean to be a pig? *The Apache Pig Project has some founding principles that help pig developers decide how the system should grow over time.*

1. **Pigs Eat Anything:** Pig can operate on data whether it has metadata or not. It can operate on data that is relational, nested, or unstructured. And it can easily be extended to operate on data beyond files, including key/value stores, databases, etc.
2. **Pigs Live Anywhere:** Pig is intended to be a language for parallel data processing. It is not tied to one particular parallel framework. It has been implemented first on Hadoop, but we do not intend that to be only on Hadoop.
3. **Pigs Are Domestic Animals:** Pig is designed to be easily controlled and modified by its users. Pig allows integration of user code where ever possible, so it currently supports user defined field transformation functions, user defined aggregates, and user defined conditionals.

   3.1 These functions can be written in Java or scripting languages that can compile down to Java (e.g. Jython). Pig supports user provided load and store functions.

Apache Pig Philosophy

What does it mean to be a pig? *The Apache Pig Project has some founding principles that help pig developers decide how the system should grow over time.*

3.2 It supports external executables via its stream command and MapReduce jars via its mapreduce command. It allows users to provide a custom partitioner for their jobs in some circumstances and to set the level of reduce parallelism for their jobs.

3.3 It allows users to set the level of reduce parallelism for their jobs and in some circumstances to provide a custom partitioner.

3.4 Pig has an optimizer that rearranges some operations in Pig Latin scripts to give better performance, combines Map Reduce jobs together, etc. However, users can easily turn this optimizer off to prevent it from making changes that do not make sense in their situation.

4. **Pigs Fly:** Pig processes data quickly. We want to consistently improve performance, and not implement features in ways that weigh pig down so it can't fly.

Apache Pig Philosophy

Pig Philosophy

Pig processes data quickly

Pigs Fly

Pigs are Domestic Animals

Pig Philosophy

Pigs Eat Anything

easily controlled and modified by its users

operate on relational, nested, or unstructured

Pigs Live Anywhere

Processes files not only from HDFS