Data Transformations with Apache Pig

INTRODUCING PIG

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Understand why Pig is used in big data analysis

Know where Pig fits in the Hadoop eco-system

Understand the differences between Pig and Hive

Introduce Pig Latin and understand how it is different from SQL and other query languages

What You Need in Your Toolkit

Prerequisites

A basic understanding of the Hadoop distributed computing framework

Familiarity with the command line on a Mac or Linux machines

Familiarity with SQL or other query languages would help, but is not necessary

Install and Setup

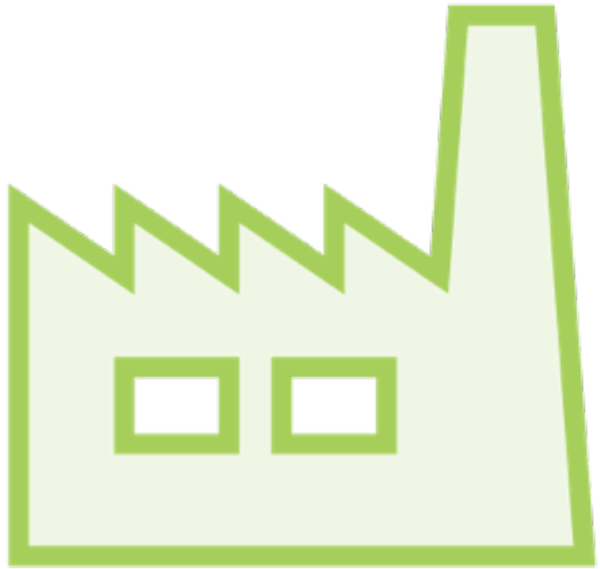
No pre-installed software needed

Basic Linux or Mac machine on which Pig can be installed

On Windows systems, use a virtual machine, Pig is commonly used on Linux

Data Drives Decisions

Organizations and Decisions



Organizations have to constantly make decisions to steer the company in the right direction

Organizations and Decisions

E-commerce site

What TV ad campaign should we run during the sale?

Compare new customer sign ups across different ad campaigns

Which one resonates the most with viewers?

Organizations and Decisions

Display ad network

Were our new format display ads more successful?

Compare the click through rates and conversions with old format

Has there been an uptick on these metrics?

Organizations and Decisions

Mall retail outlets

Where should we set up our store in the new mall?

Compare foot traffic from store sensors at existing locations

What location sees the most footfalls and conversions?

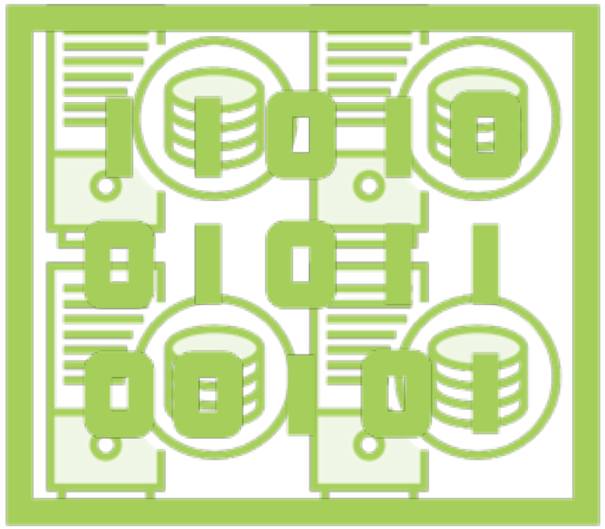
Decisions Require Data

show current state of affairs

indicate trends

predict behaviors

Decisions Require Data



Data for analytical processing is typically stored in a data warehouse

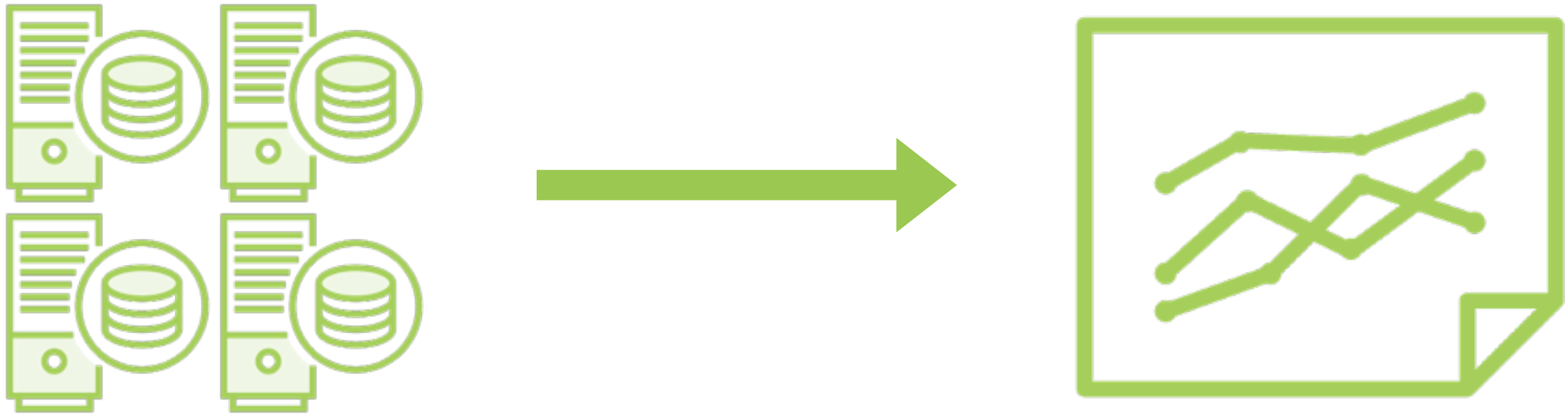
Data Warehouse

huge dataset from multiple sources

semi-structured data

long running jobs to extract information

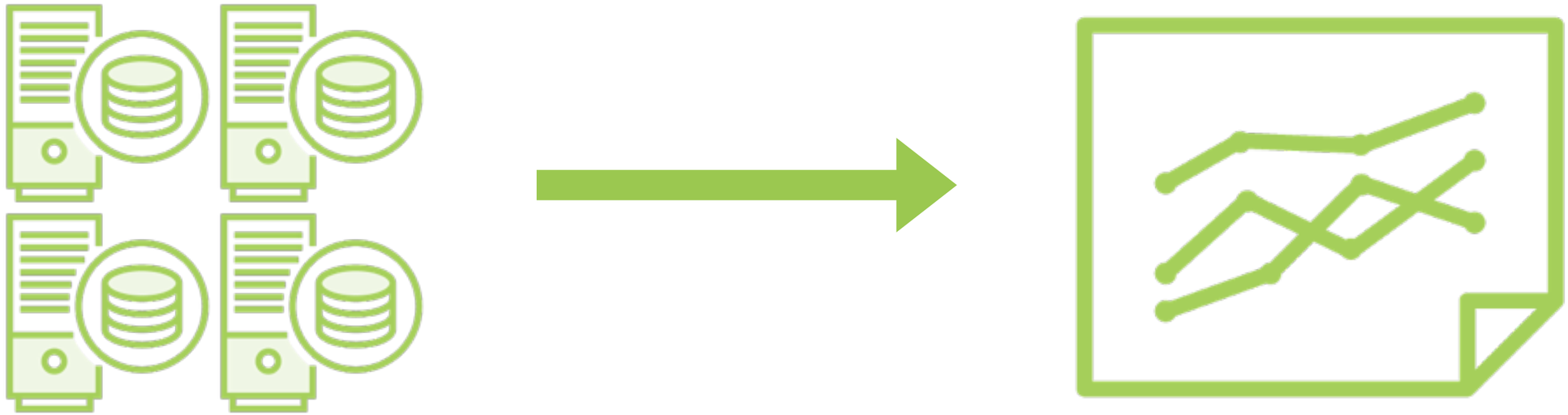
Data Warehouse



Can be used to extract meaningful information which drives decisions

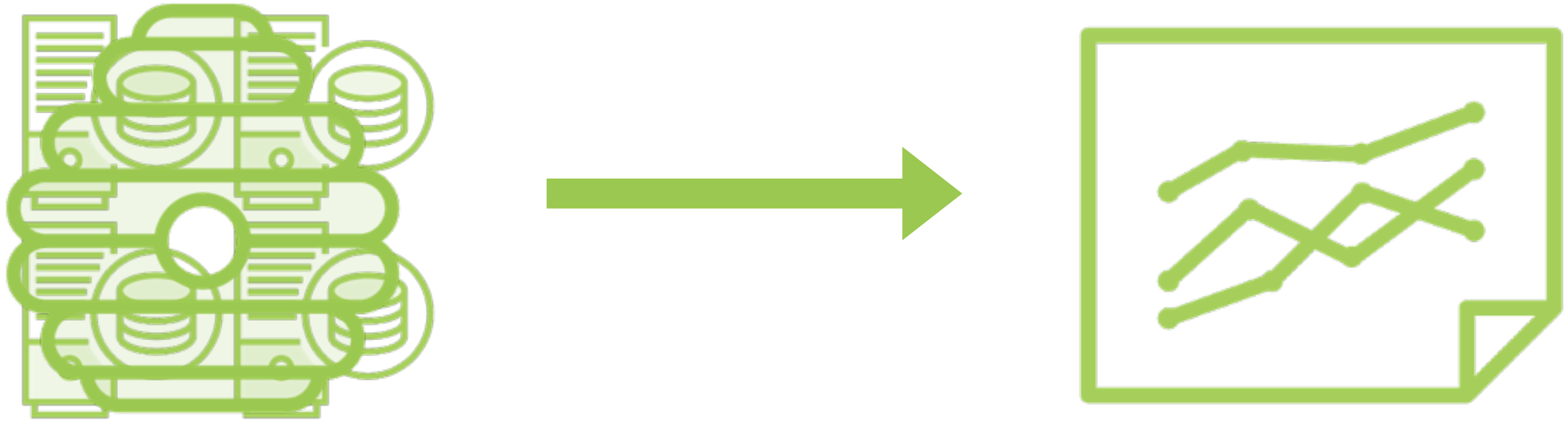
Hive as a Data Warehouse

Data Warehouse



A data warehouse stores data that is processed for insights

Data Warehouse



Apache Hive is an open-source data warehouse

HiveQL to query data

Hadoop

HDFS

MapReduce

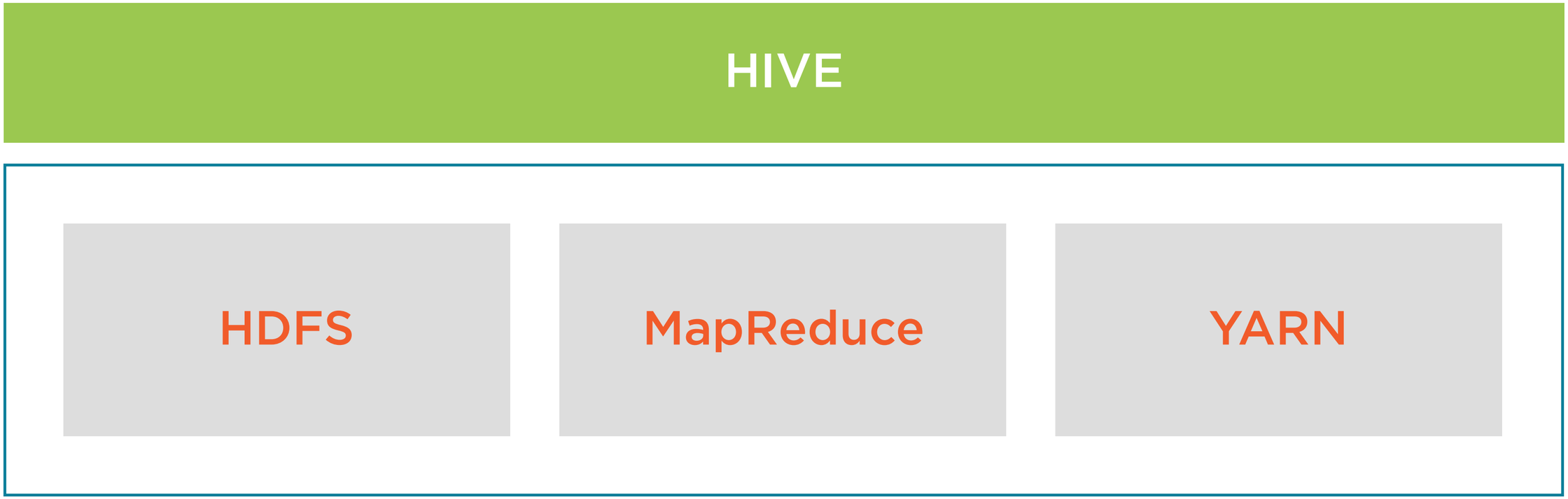
YARN

File system to manage the storage of data

Framework to process data across multiple servers

Framework to run and manage the data processing tasks

Hive on Hadoop

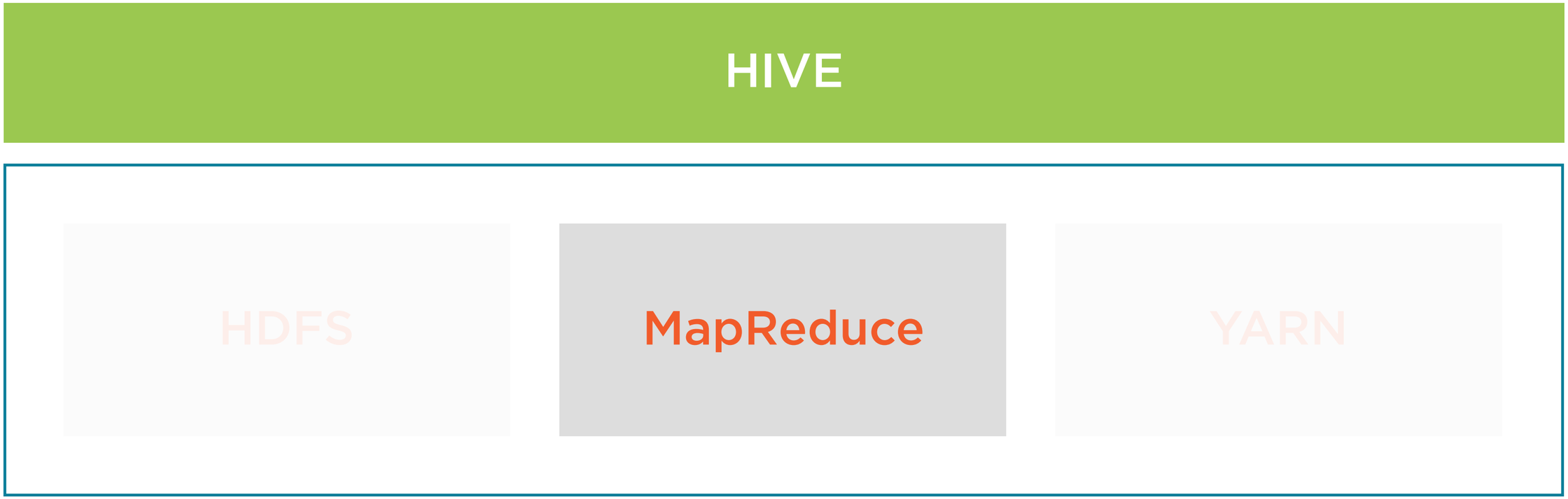


Hive runs on top of the Hadoop distributed computing framework

Hive on Hadoop

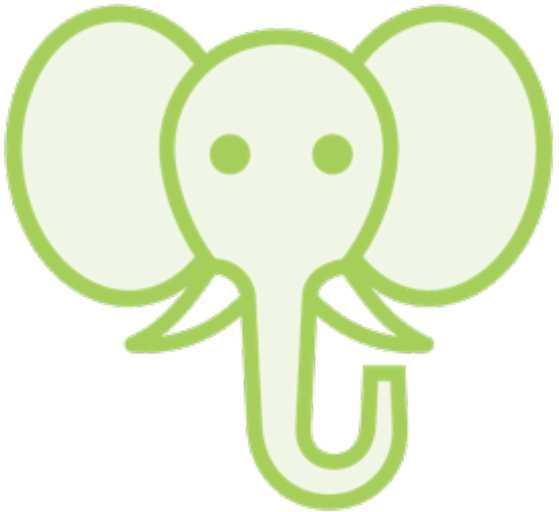
Hive stores its data in HDFS

Hive on Hadoop



Hive runs all processes in the form of MapReduce jobs under the hood

Hive on Hadoop



Allows processing of huge datasets

Errr... okay....

Where Does Data Come From?

Data Sources and Characteristics

Where Does Data Come From?

Where Does Data Come From?

Mobile phones

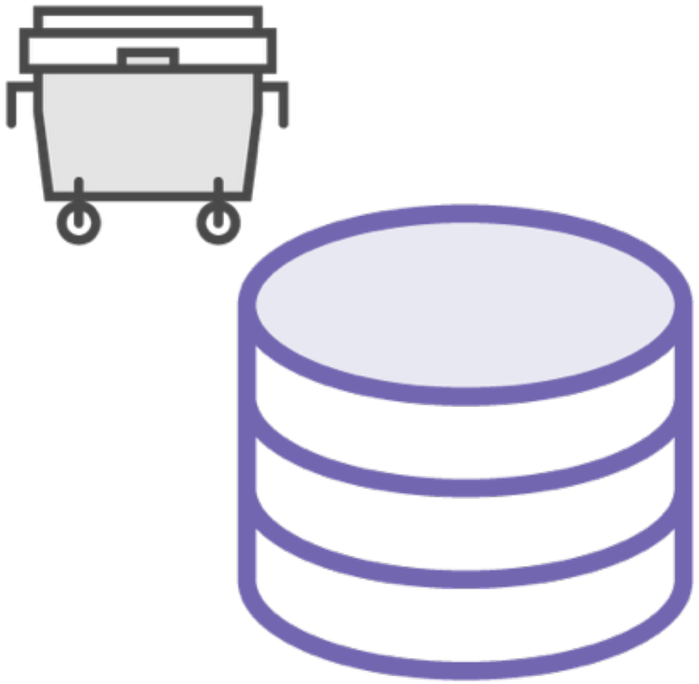
GPS location coordinates

Self driving car sensors

Where Is This Data Stored?

Files on some storage system

Characteristics of Data

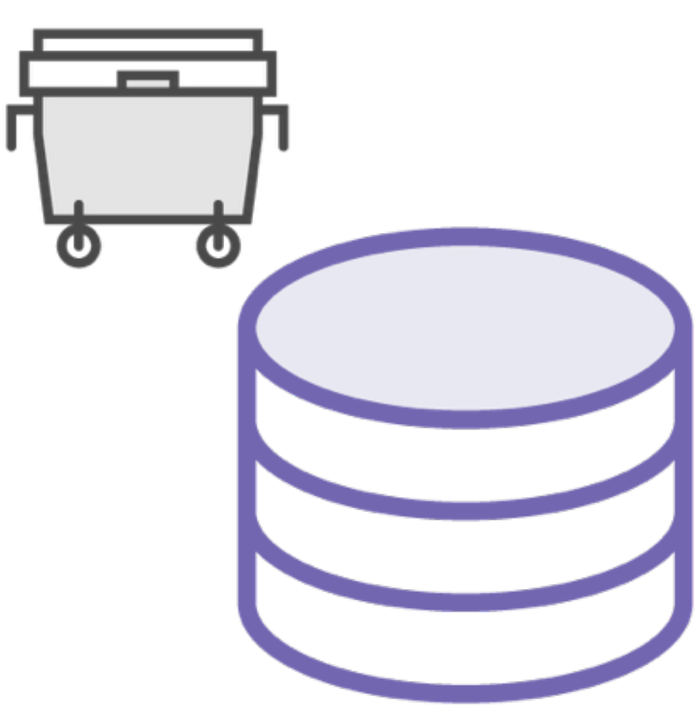


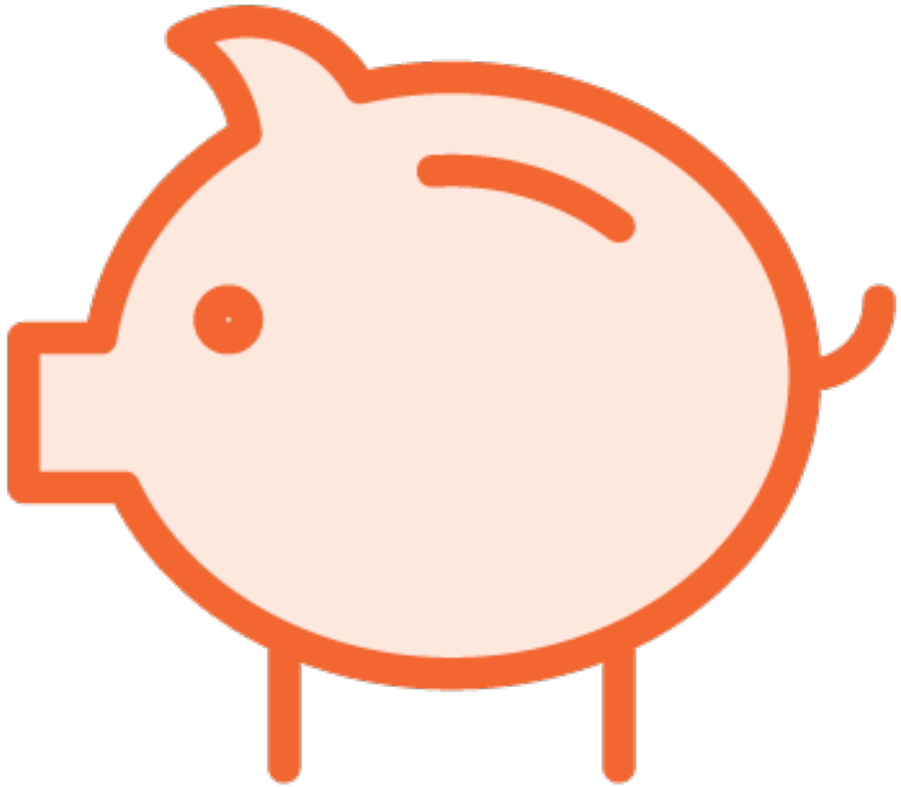
Unknown schema

Incomplete data

Inconsistent records

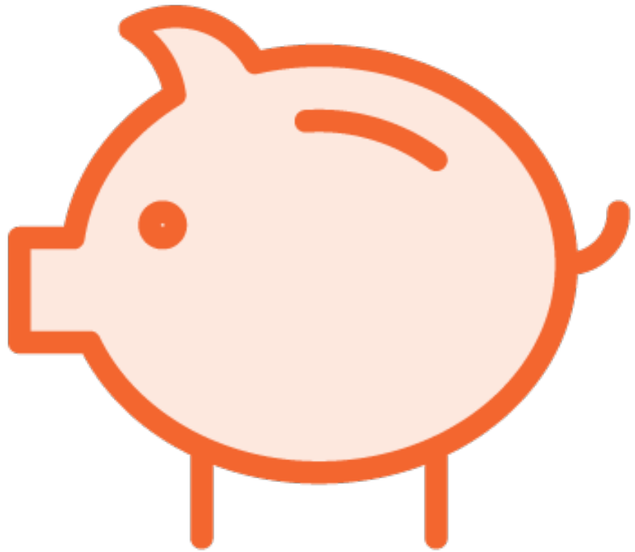
Apache Pig to the Rescue



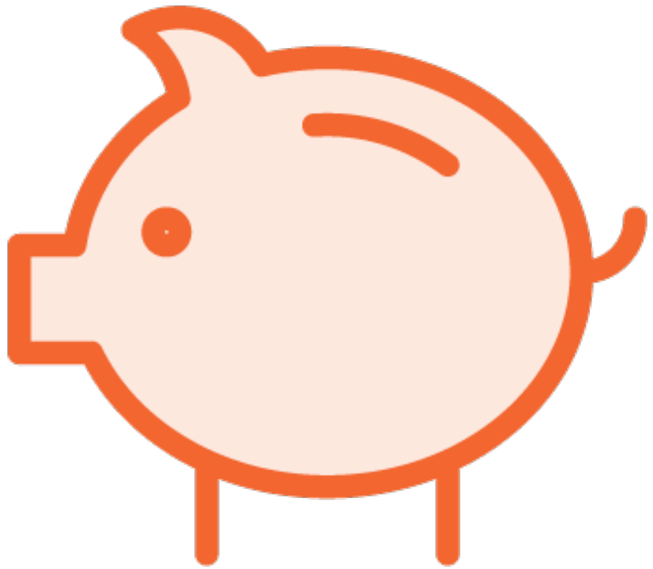


How Can Pig Help Us?

Apache Pig to the Rescue



A high level scripting language to work with data with unknown or inconsistent schema



Pig

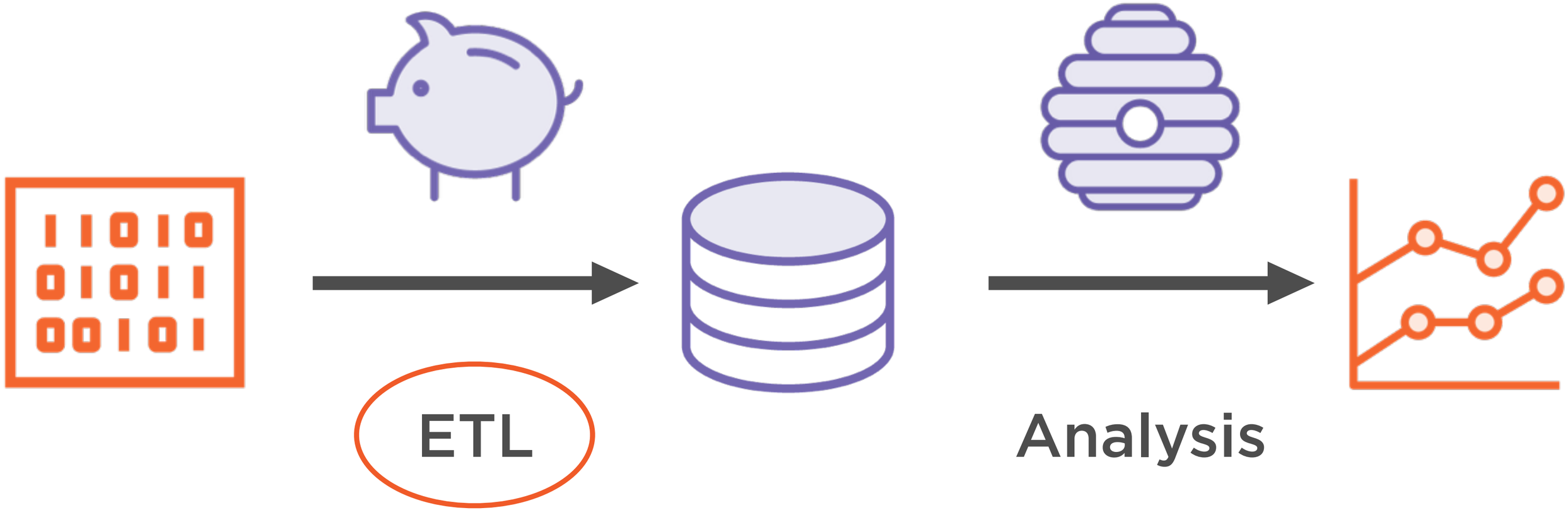
Part of the Hadoop eco-system

Works well with unstructured, incomplete data

Can work directly on files in HDFS

Used to get data into a data warehouse

Pig Complements Hive



Extract, Transform, Load

Pull unstructured, inconsistent data from source, clean it and place it in another database where it can be analyzed

Extract, Transform, Load

Pull unstructured, inconsistent data from source, clean it and place it in another database where it can be analyzed

Extract, Transform, Load

Pull unstructured, inconsistent data from source, clean it and place it in another database where it can be analyzed

Extract, Transform, Load

Pull unstructured, inconsistent data from source, clean it and **place it in another database where it can be analyzed**

Apache Pig

```
64.242.88.10 - - [07/Mar/2004:16:05:49 -0800] "GET /twiki/bin/edit/Main/Double_bounce_sender?topicparent=Main.ConfigurationVariables
HTTP/1.1" 401 12846
64.242.88.10 - - [07/Mar/2004:16:06:51 -0800] "GET /twiki/bin/rdiff/TWiki/NewUserTemplate?rev1=1.3&rev2=1.2 HTTP/1.1" 200 4523
64.242.88.10 - - [07/Mar/2004:16:10:02 -0800] "GET /mailman/listinfo/hsdivision HTTP/1.1" 200 6291
64.242.88.10 - - [07/Mar/2004:16:11:58 -0800] "GET /twiki/bin/view/TWiki/WikiSyntax HTTP/1.1" 200 7352
64.242.88.10 - - [07/Mar/2004:16:20:55 -0800] "GET /twiki/bin/view/Main/DCCAndPostFix HTTP/1.1" 200 5253
64.242.88.10 - - [07/Mar/2004:16:23:12 -0800] "GET /twiki/bin/oops/TWiki/AppendixFileSystem?template=oopsmore¶m1=1.12¶m2=1.12 HTTP/1.1"
200 11382
64.242.88.10 - - [07/Mar/2004:16:24:16 -0800] "GET /twiki/bin/view/Main/PeterThoeny HTTP/1.1" 200 4924
64.242.88.10 - - [07/Mar/2004:16:29:16 -0800] "GET /twiki/bin/edit/Main/Header_checks?topicparent=Main.ConfigurationVariables HTTP/1.1"
401 12851
64.242.88.10 - - [07/Mar/2004:16:30:29 -0800] "GET /twiki/bin/attach/Main/OfficeLocations HTTP/1.1" 401 12851
64.242.88.10 - - [07/Mar/2004:16:31:48 -0800] "GET /twiki/bin/view/TWiki/WebTopicEditTemplate HTTP/1.1" 200 3732
64.242.88.10 - - [07/Mar/2004:16:32:50 -0800] "GET /twiki/bin/view/Main/WebChanges HTTP/1.1" 200 40520
64.242.88.10 - - [07/Mar/2004:16:33:53 -0800] "GET /twiki/bin/edit/Main/Smtpd_etrn_restrictions?topicparent=Main.ConfigurationVariables
HTTP/1.1" 401 12851
64.242.88.10 - - [07/Mar/2004:16:35:19 -0800] "GET /mailman/listinfo/business HTTP/1.1" 200 6379
64.242.88.10 - - [07/Mar/2004:16:36:22 -0800] "GET /twiki/bin/rdiff/Main/WebIndex?rev1=1.2&rev2=1.1 HTTP/1.1" 200 46373
64.242.88.10 - - [07/Mar/2004:16:37:27 -0800] "GET /twiki/bin/view/TWiki/DontNotify HTTP/1.1" 200 4140
64.242.88.10 - - [07/Mar/2004:16:39:24 -0800] "GET /twiki/bin/view/Main/TokyoOffice HTTP/1.1" 200 3853
64.242.88.10 - - [07/Mar/2004:16:43:54 -0800] "GET /twiki/bin/view/Main/MikeMannix HTTP/1.1" 200 3686
64.242.88.10 - - [07/Mar/2004:16:45:56 -0800] "GET /twiki/bin/attach/Main/PostfixCommands HTTP/1.1" 401 12846
64.242.88.10 - - [07/Mar/2004:16:47:12 -0800] "GET /robots.txt HTTP/1.1" 200 68
64.242.88.10 - - [07/Mar/2004:16:47:46 -0800] "GET /twiki/bin/rdiff/Know/ReadmeFirst?rev1=1.5&rev2=1.4 HTTP/1.1" 200 5724
64.242.88.10 - - [07/Mar/2004:16:49:04 -0800] "GET /twiki/bin/view/Main/TWikiGroups?rev=1.2 HTTP/1.1" 200 5162
64.242.88.10 - - [07/Mar/2004:16:50:54 -0800] "GET /twiki/bin/rdiff/Main/ConfigurationVariables HTTP/1.1" 200 59679
64.242.88.10 - - [07/Mar/2004:16:52:35 -0800] "GET /twiki/bin/edit/Main/Flush_service_name?topicparent=Main.ConfigurationVariables
HTTP/1.1" 401 12851
```

Apache Pig

```
64.242.88.10 - - [07/Mar/2004:16:05:49 -0800] "GET /twiki/bin/
edit/Main/Double_bounce_sender?
topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12846
```

Server IP Address

Apache Pig

```
64.242.88.10 - - [07/Mar/2004:16:05:49 -0800] "GET /twiki/bin/
edit/Main/Double_bounce_sender?
topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12846
```

Date and Time

Apache Pig

```
64.242.88.10 - - [07/Mar/2004:16:05:49 -0800] "GET /twiki/bin/
edit/Main/Double_bounce_sender?
topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12846
```

Request Type

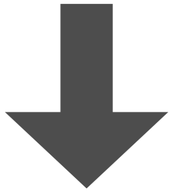
Apache Pig

```
64.242.88.10 - - [07/Mar/2004:16:05:49 -0800] "GET /twiki/bin/
edit/Main/Double_bounce_sender?
topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12846
```

URL

Apache Pig

```
64.242.88.10 - - [07/Mar/2004:16:05:49 -0800] "GET /twiki/bin/
edit/Main/Double_bounce_sender?
topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12846
```



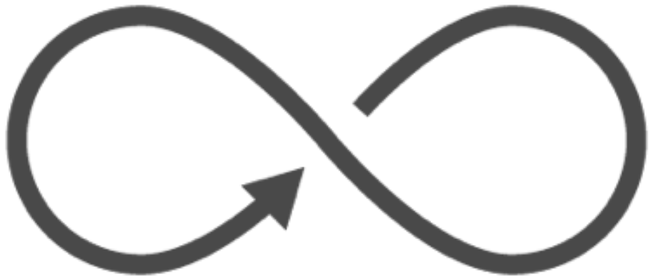
IP	Date	Time	Request Type	URL

Pig Latin

Pig Latin

A procedural, data flow language to extract, transform and load data

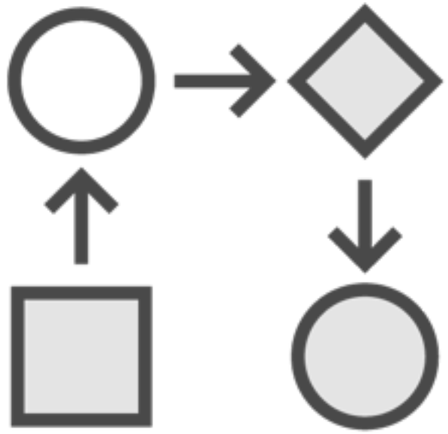
Procedural



Series of well-defined steps to perform operations

No if statements or for loops

Data Flow



Focused on transformations applied to the data

Written with a series of data operations in mind

Pig Latin

Data from one or more sources can be read, processed and stored in parallel

Pig Latin

Cleans data, precomputes common aggregates before storing in a data warehouse

Pig vs. SQL

Pig

```
(group revenues by dept)
                foreach
              generate
           sum(revenue)
```

SQL

```
select sum(revenue)
from revenues
group by dept
```

Pig vs. SQL

Pig	SQL
A data flow language, transforms data to store in a warehouse	A query language, is used for retrieving results
Specifies exactly how data is to be modified at every step	Abstracts away how queries are executed
Purpose of processing is to store in a queryable format	Purpose of data extraction is analysis
Used to clean data with inconsistent or incomplete schema	Extract insights, generate reports, drive decisions

Pig on Hadoop

Hadoop

HDFS

MapReduce

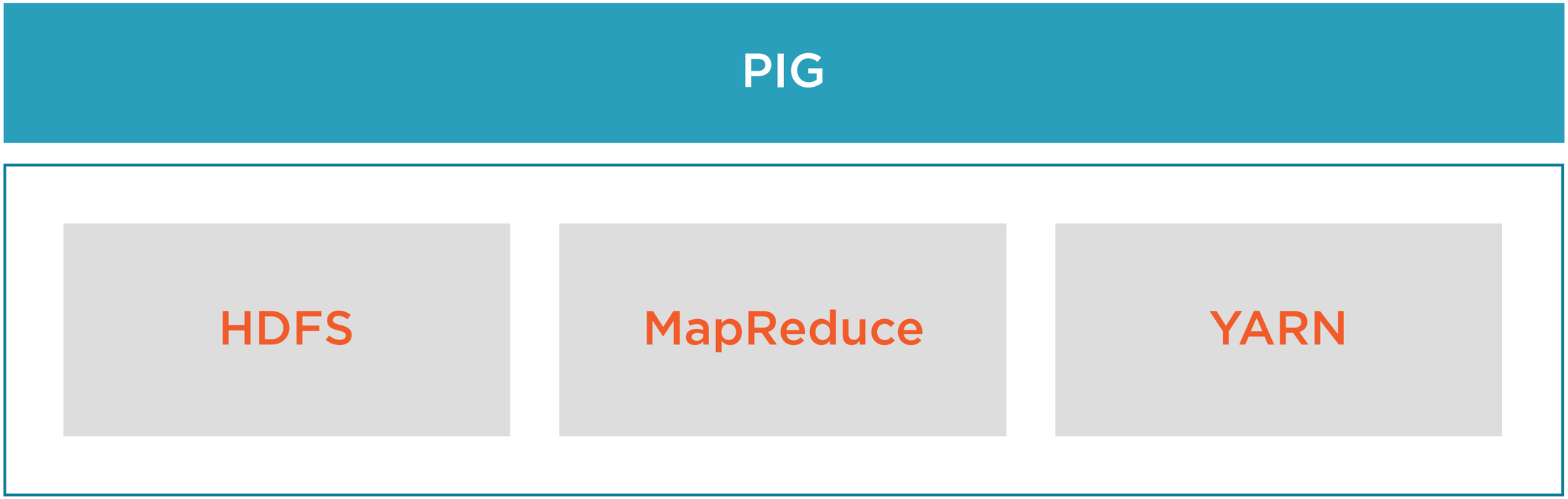
YARN

File system to manage the storage of data

Framework to process data across multiple servers

Framework to run and manage the data processing tasks

Pig on Hadoop

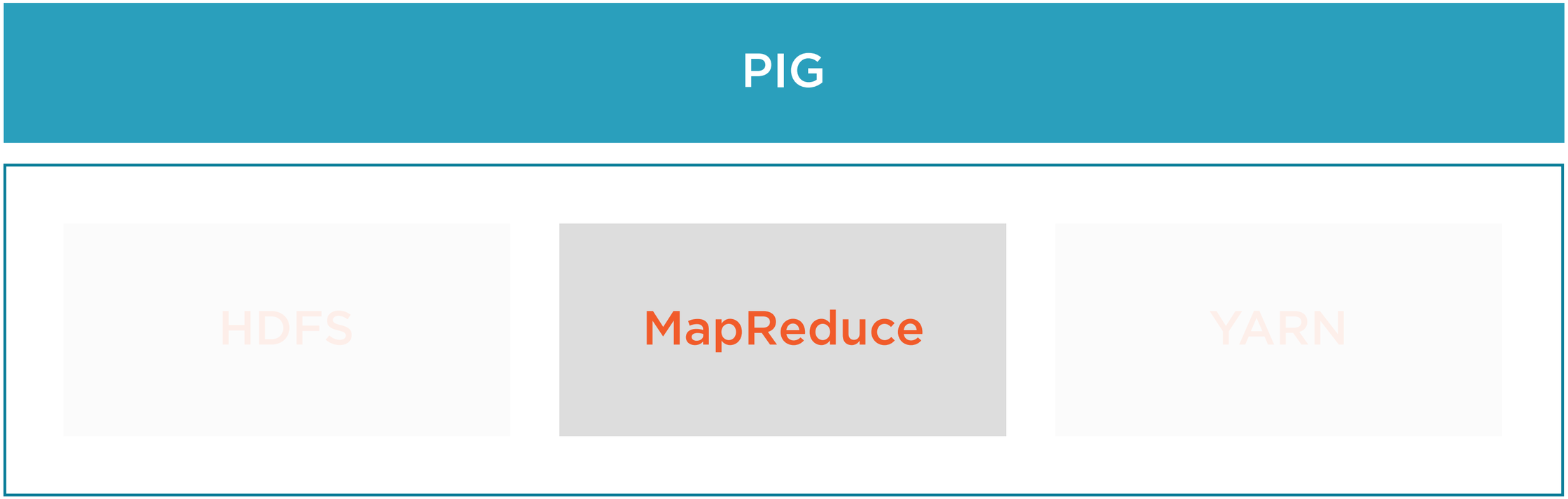


Pig runs on top of the Hadoop distributed computing framework

Pig on Hadoop

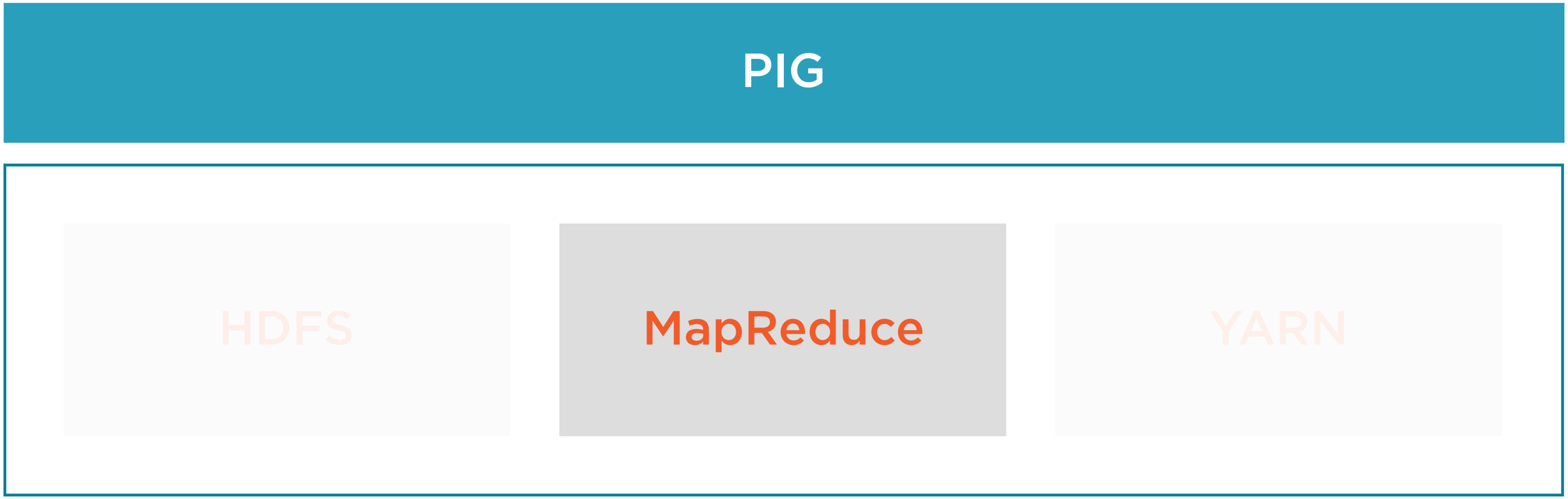
Reads files from HDFS, stores intermediate records in HDFS and writes its final output to HDFS

Pig on Hadoop



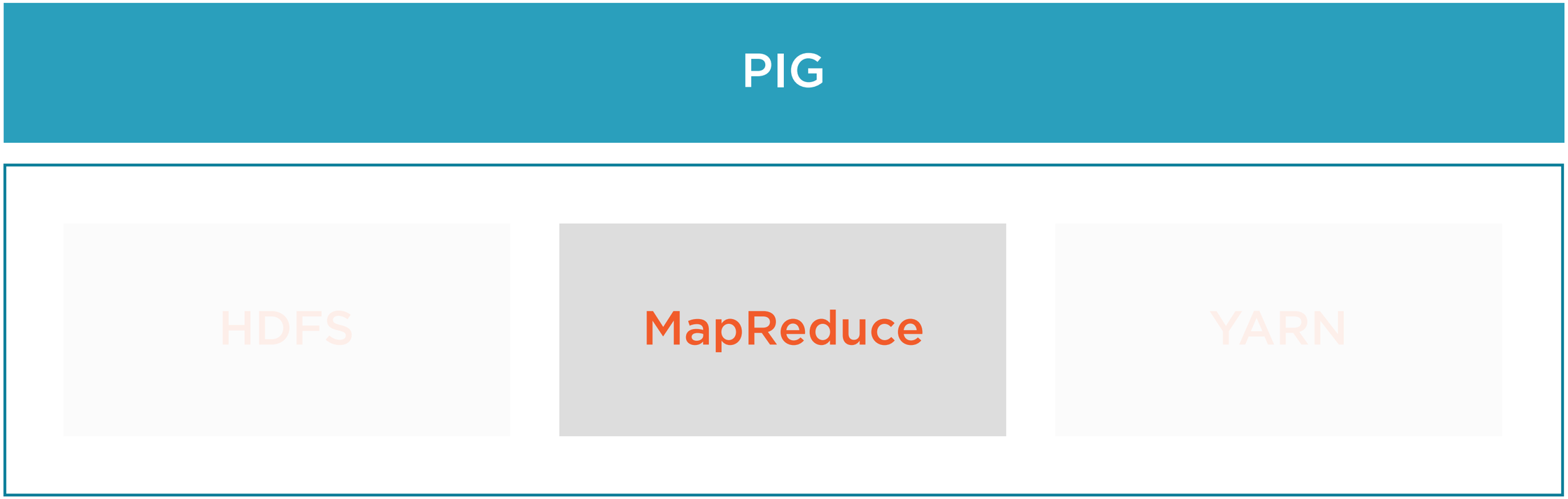
Decomposes operations into MapReduce jobs which run in parallel

Pig on Hadoop



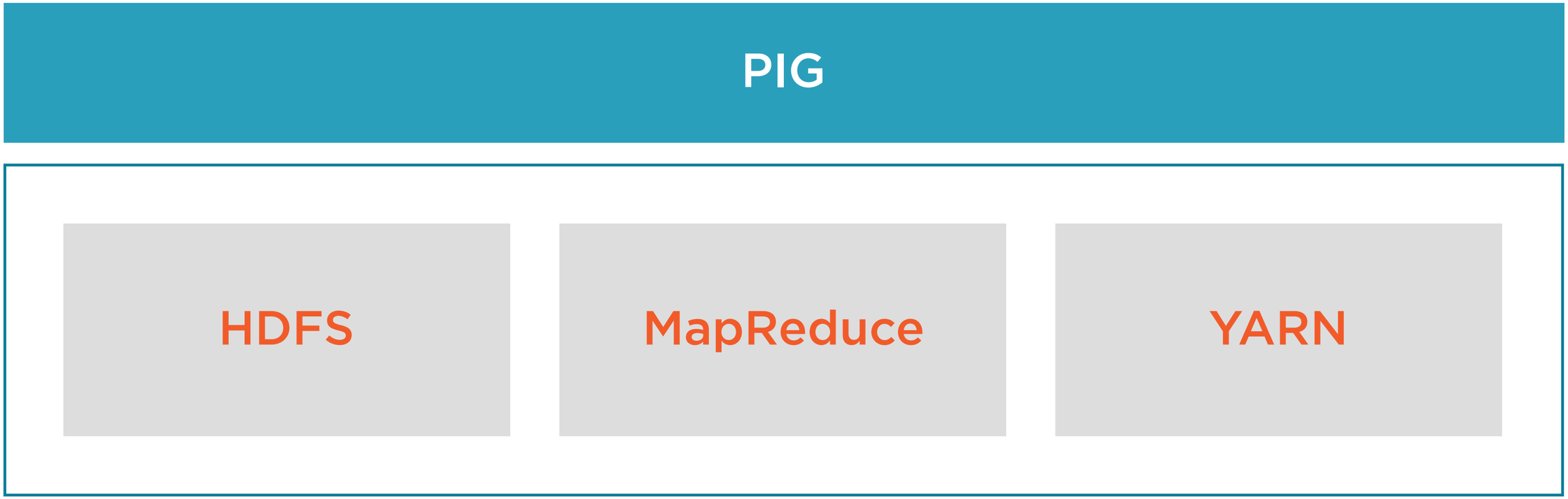
Provides non-trivial, built-in implementations of standard data operations, which are very efficient

Pig on Hadoop

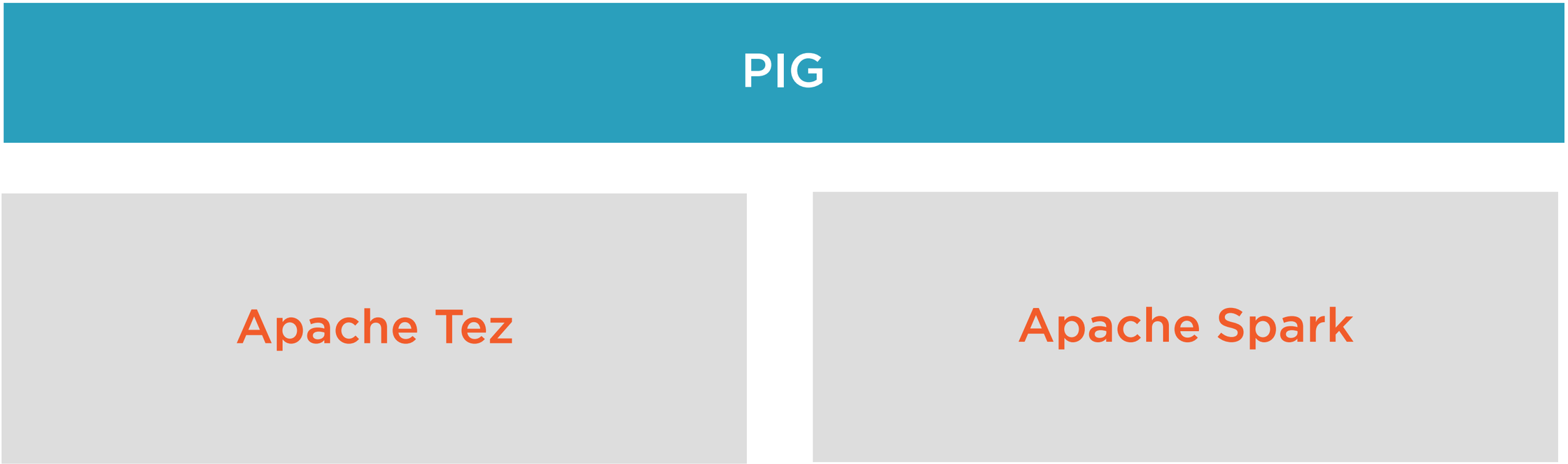


Pig optimizes operations before MapReduce jobs are run, to speed operations up

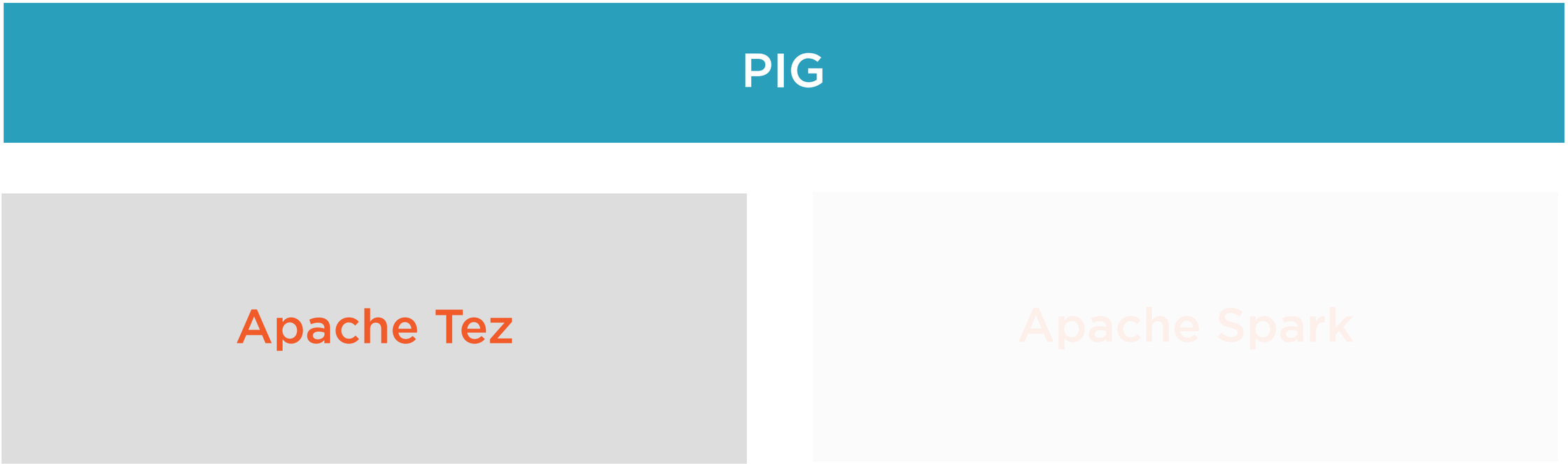
Pig on Hadoop



Pig on Other Technologies



Pig on Other Technologies



Tez is an extensible framework which improves on MapReduce by making its operations faster

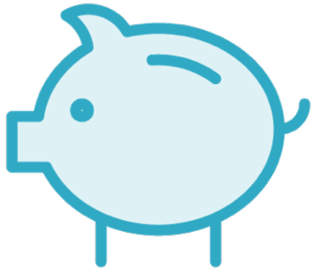
Pig on Other Technologies

PIG

Apache Tez

Apache Spark

Spark is another distributed computing technology which is scalable, flexible and fast



Pig vs. Hive

Pig	Hive
Used to extract, transform and load data into a data warehouse	Used to query data from a data warehouse to generate reports
Used by developers to bring together useful data in one place	Used by analysts to retrieve business information from data
Uses Pig Latin, a procedural, data flow language	Uses HiveQL, a structured query language

Summary

Understood the importance of Pig to extract, transform and load data

Know the role of Pig in the Hadoop eco-system and how it complements the working of Hive

Understood the use of Pig Latin and how it differs from SQL