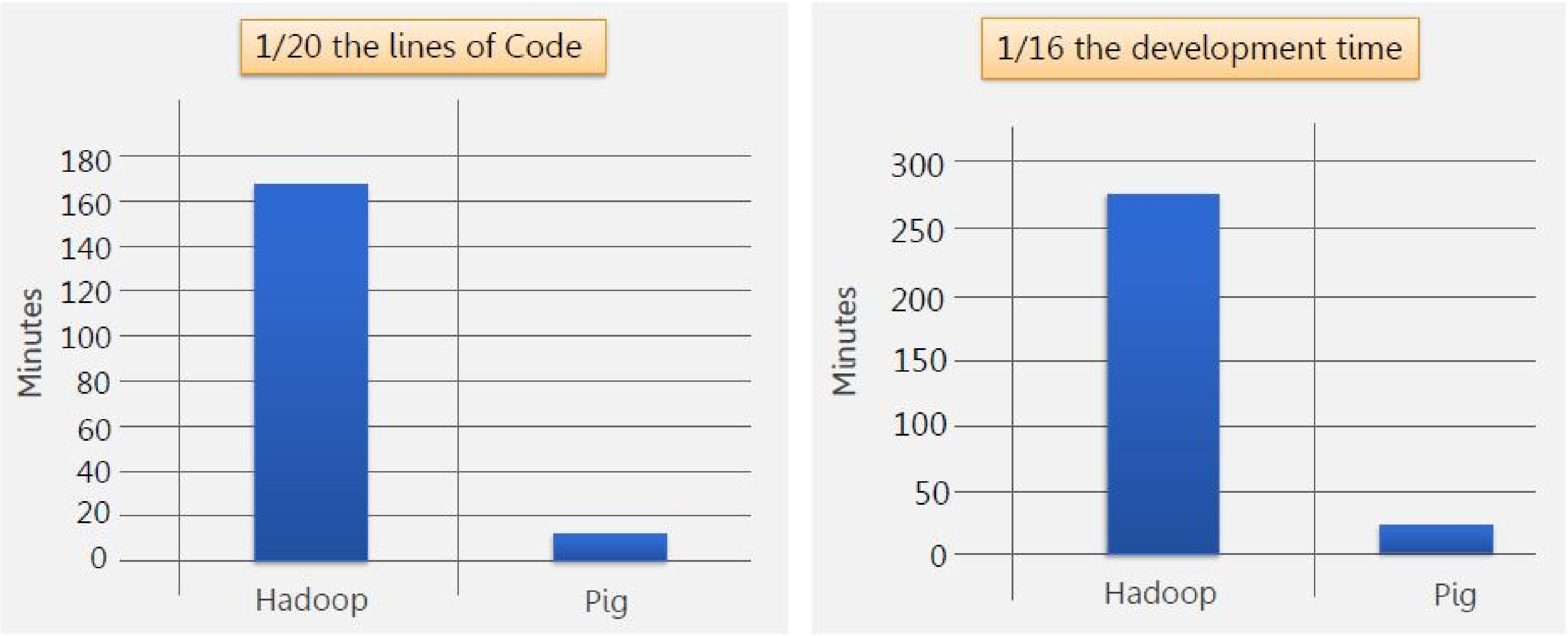
PIG

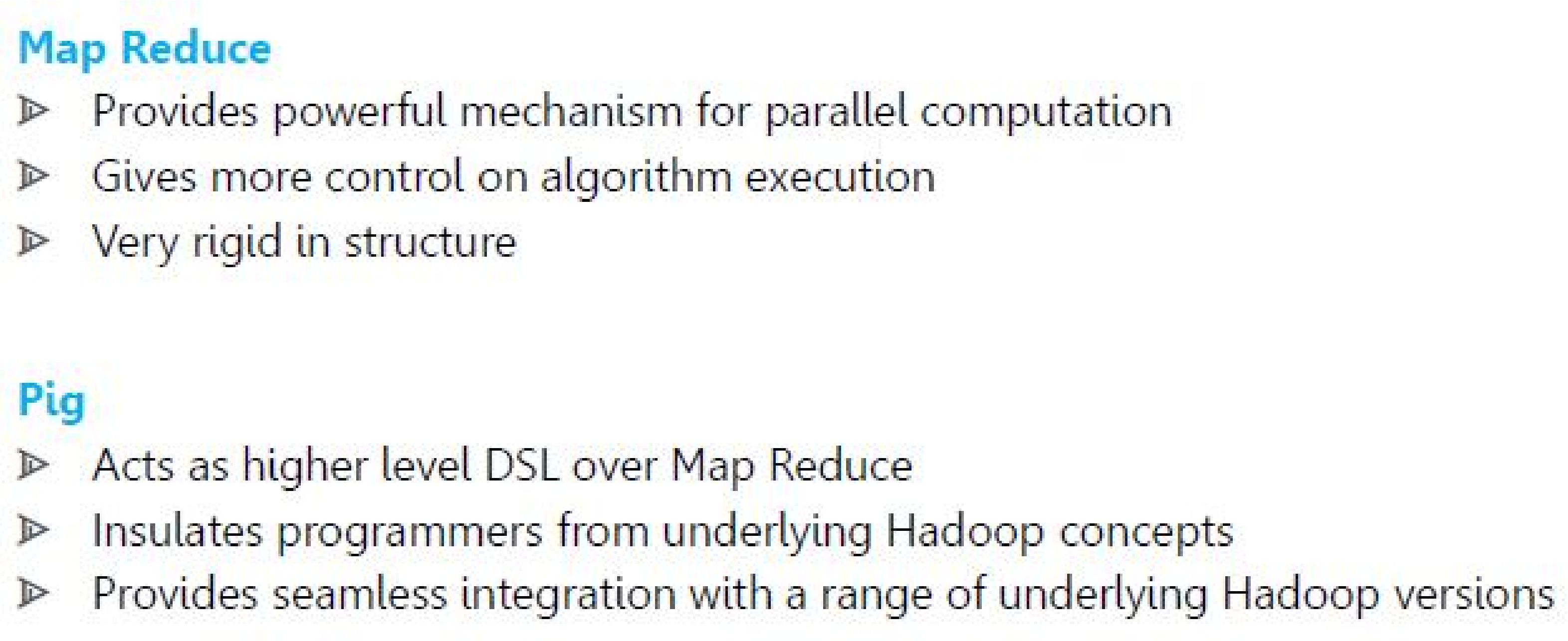
Outline

- Introduction to Pig
- MR vs Pig
- Pig Architecture
- Pig Data Types
- Pig Operators
- Pig Built-in Functions

MR vs Pig


1/20 the lines of Code
Minutes
180
160
140
120
100
80
60
40
20
0
Hadoop
Pig
1/16 the development time
Minutes
300
250
200
150
100
50
0
Hadoop
Pig

MR vs Pig



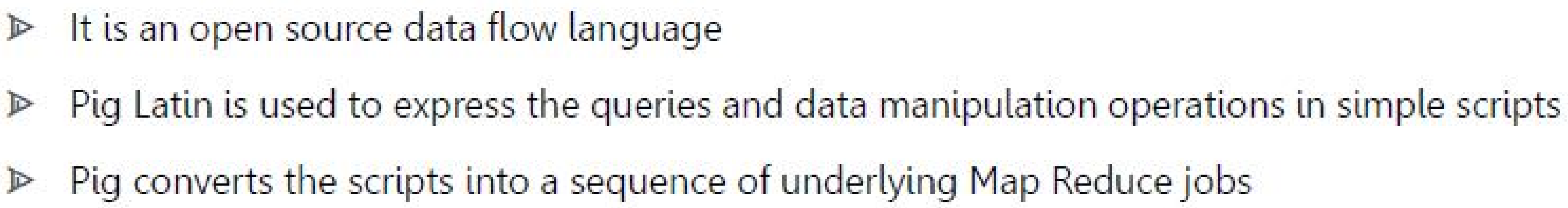
Map Reduce

- Provides powerful mechanism for parallel computation
- Gives more control on algorithm execution
- Very rigid in structure

Pig

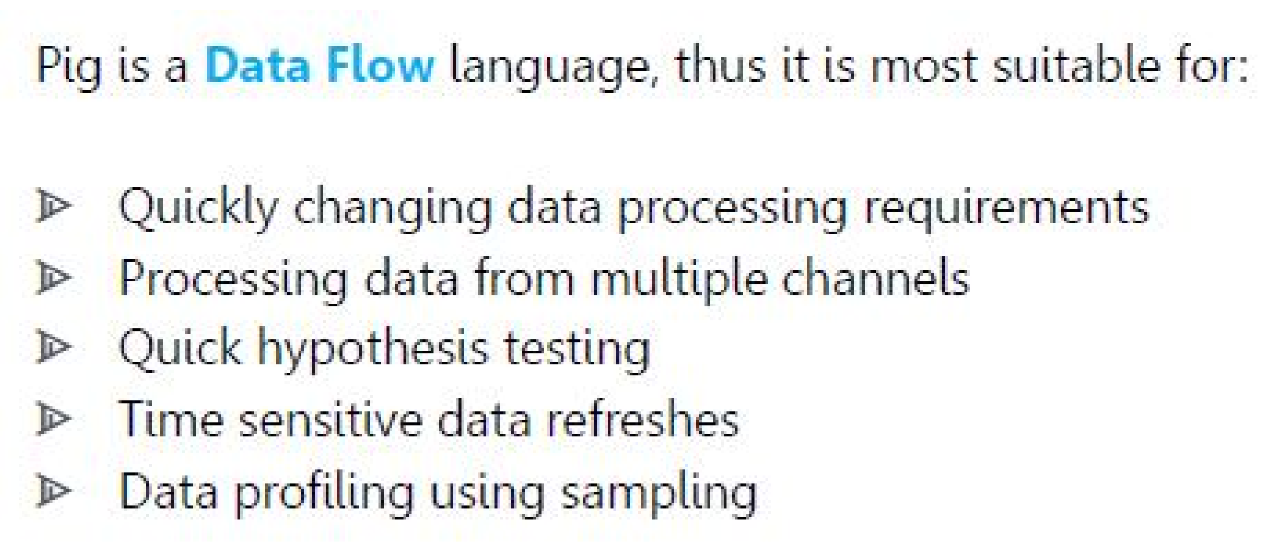
- Acts as higher level DSL over Map Reduce
- Insulates programmers from underlying Hadoop concepts
- Provides seamless integration with a range of underlying Hadoop versions

What is Pig?



- It is an open source data flow language
- Pig Latin is used to express the queries and data manipulation operations in simple scripts
- Pig converts the scripts into a sequence of underlying Map Reduce jobs

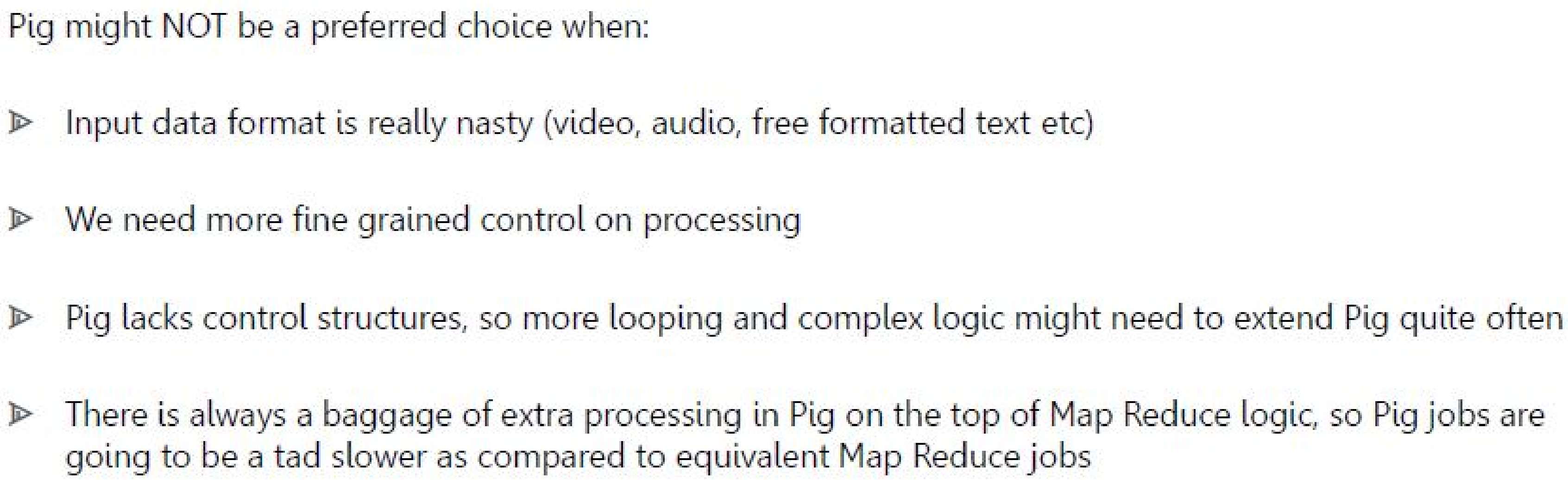
Where to use Pig?



Pig is a **Data Flow** language, thus it is most suitable for:

- Quickly changing data processing requirements
- Processing data from multiple channels
- Quick hypothesis testing
- Time sensitive data refreshes
- Data profiling using sampling

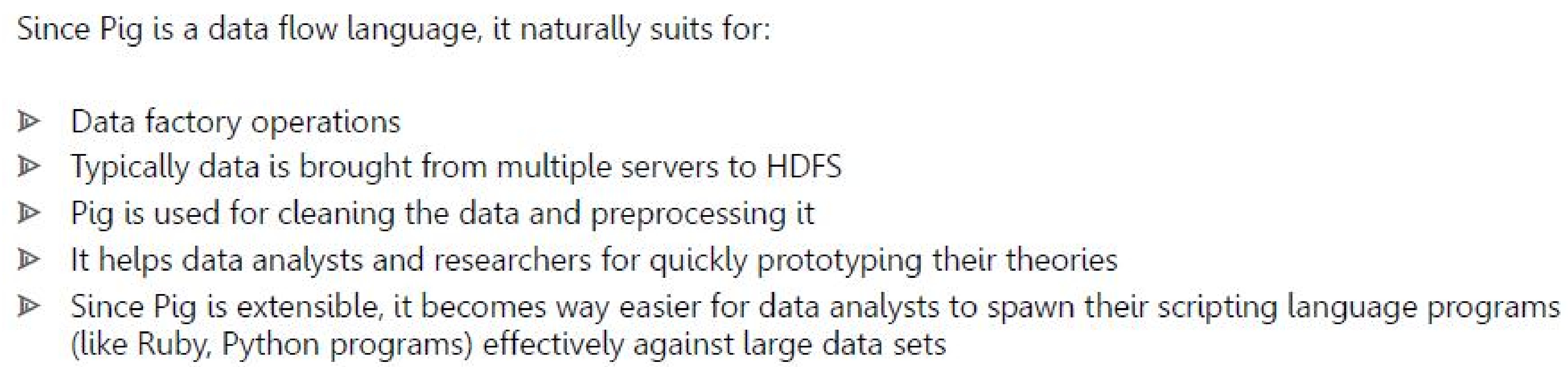
Where NOT to use Pig?



Pig might NOT be a preferred choice when:

- Input data format is really nasty (video, audio, free formatted text etc)
- We need more fine grained control on processing
- Pig lacks control structures, so more looping and complex logic might need to extend Pig quite often
- There is always a baggage of extra processing in Pig on the top of Map Reduce logic, so Pig jobs are going to be a tad slower as compared to equivalent Map Reduce jobs

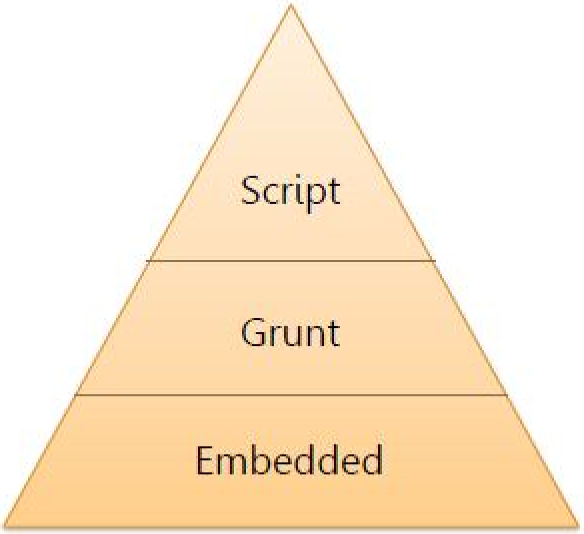
Pig in Industry



Since Pig is a data flow language, it naturally suits for:

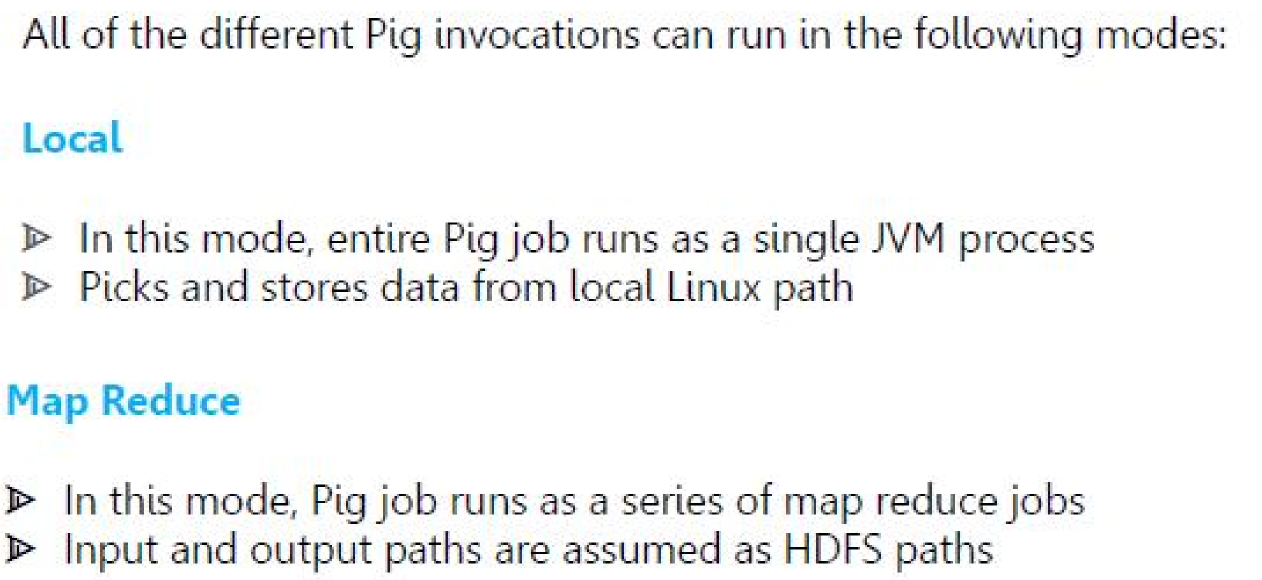
- Data factory operations
- Typically data is brought from multiple servers to HDFS
- Pig is used for cleaning the data and preprocessing it
- It helps data analysts and researchers for quickly prototyping their theories
- Since Pig is extensible, it becomes way easier for data analysts to spawn their scripting language programs (like Ruby, Python programs) effectively against large data sets

Ways to handle Pig



- Grunt Mode:
  - It's interactive mode of Pig
  - Very useful for testing syntax checking and ad-hoc data exploration
- Script Mode:
  - Runs set of instructions from a file
  - Similar to a SQL script file
- Embedded Mode:
  - Executes Pig programs from a Java program
  - Suitable to create Pig Scripts on the fly

Modes of Pig



All of the different Pig invocations can run in the following modes:

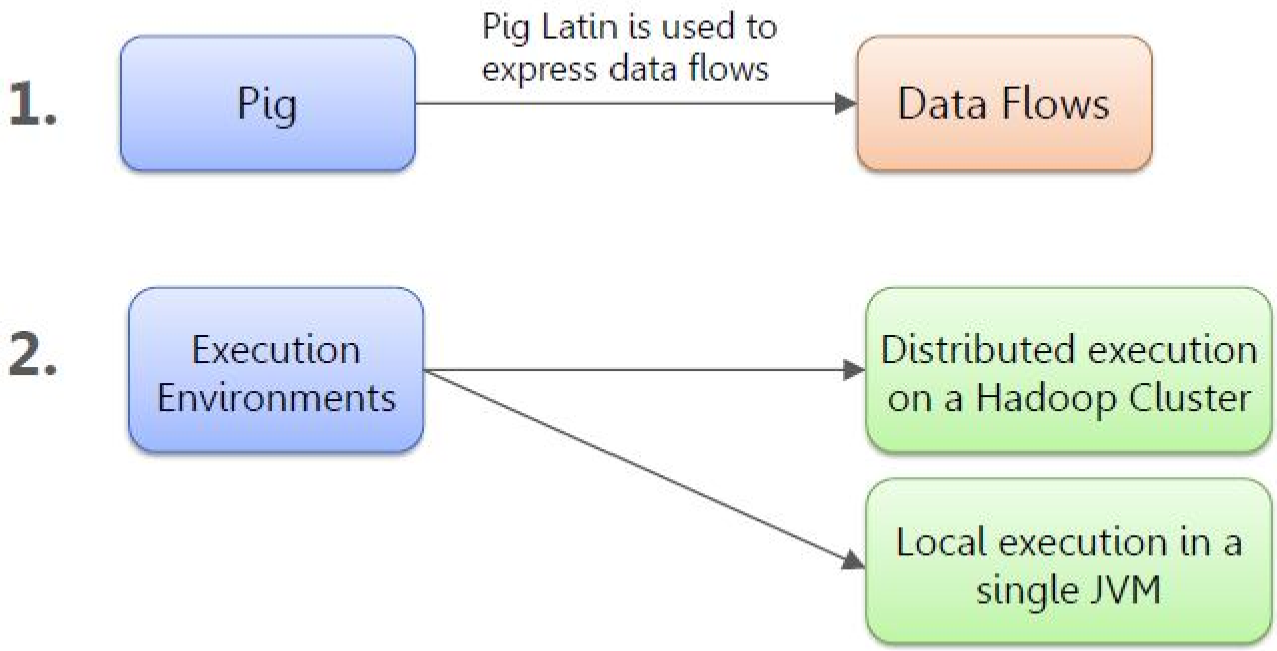
Local

- In this mode, entire Pig job runs as a single JVM process
- Picks and stores data from local Linux path

Map Reduce

- In this mode, Pig job runs as a series of map reduce jobs
- Input and output paths are assumed as HDFS paths

Pig Components



1. Pig

Pig Latin is used to express data flows

Data Flows

2. Execution Environments

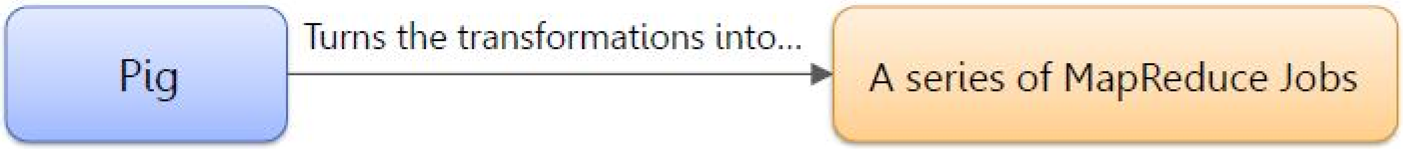
Distributed execution on a Hadoop Cluster

Local execution in a single JVM

Pig Programs Execution

Pig is just a wrapper on top of Map Reduce layer

It parses, optimizes and converts the Pig script to a series of Map Reduce jobs

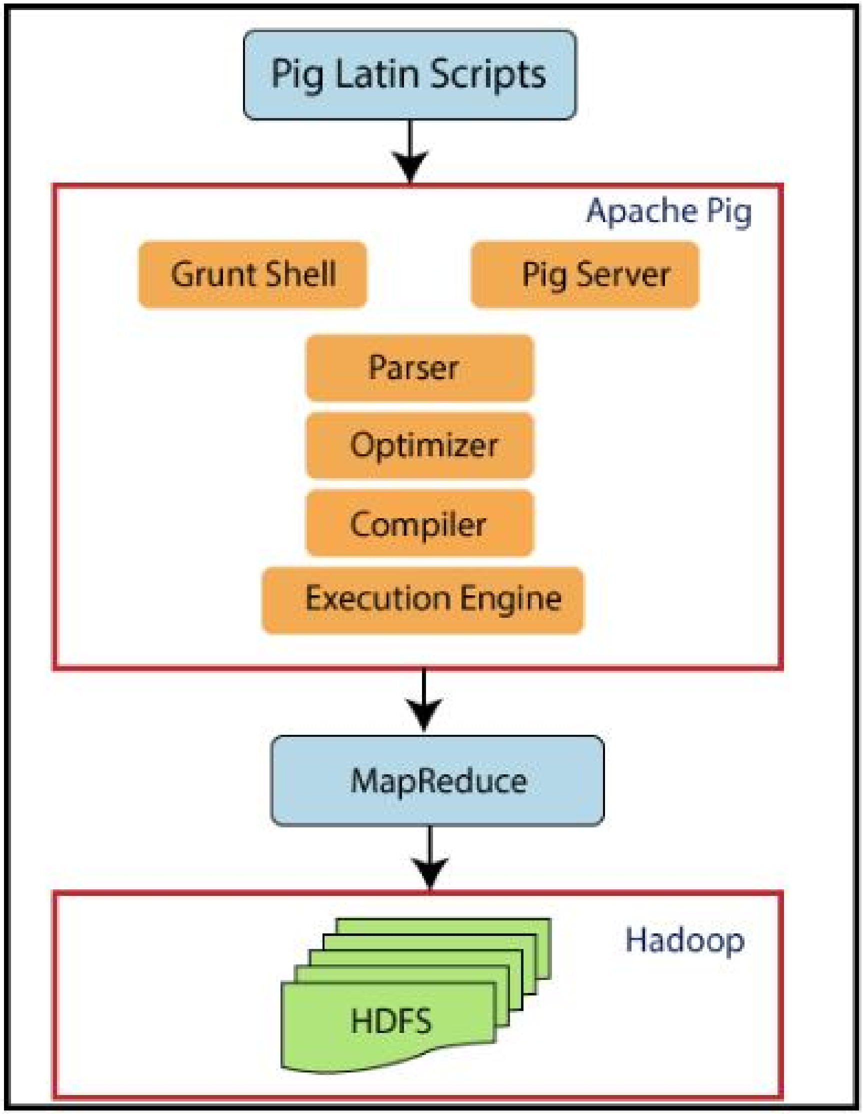


Pig

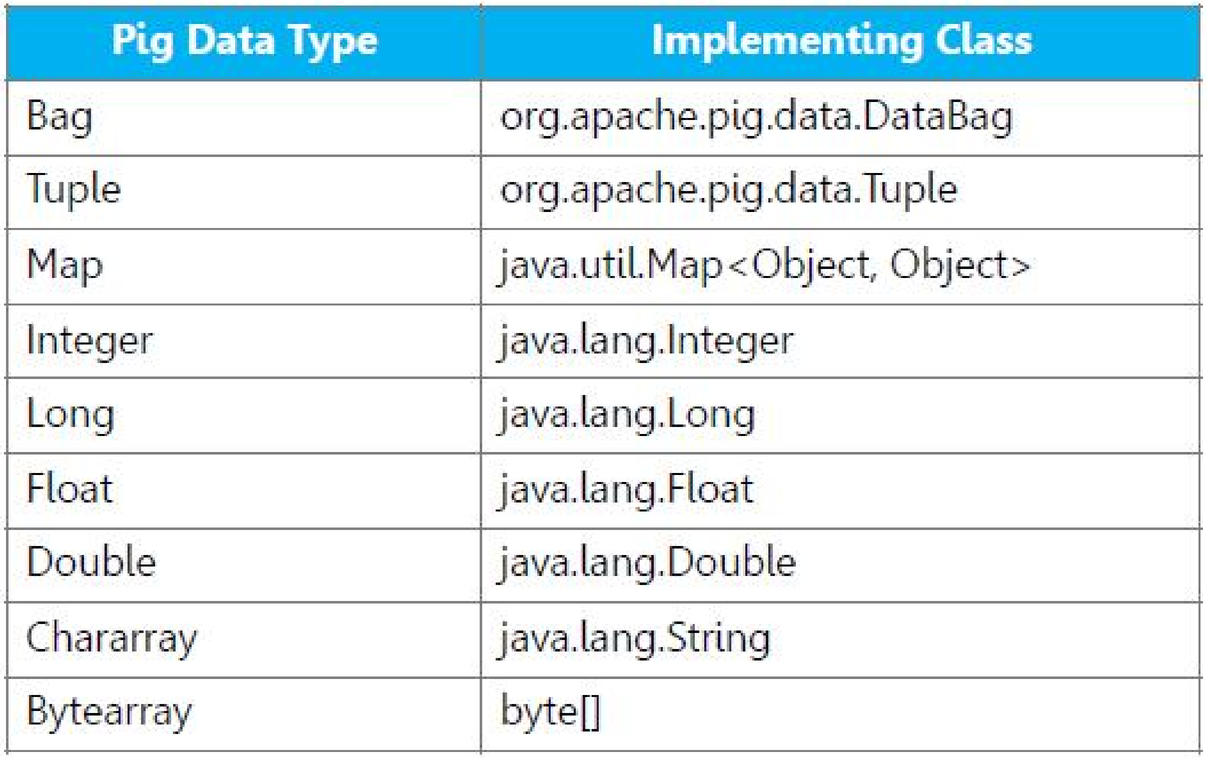
Turns the transformations into...

A series of MapReduce Jobs

Pig Architecture

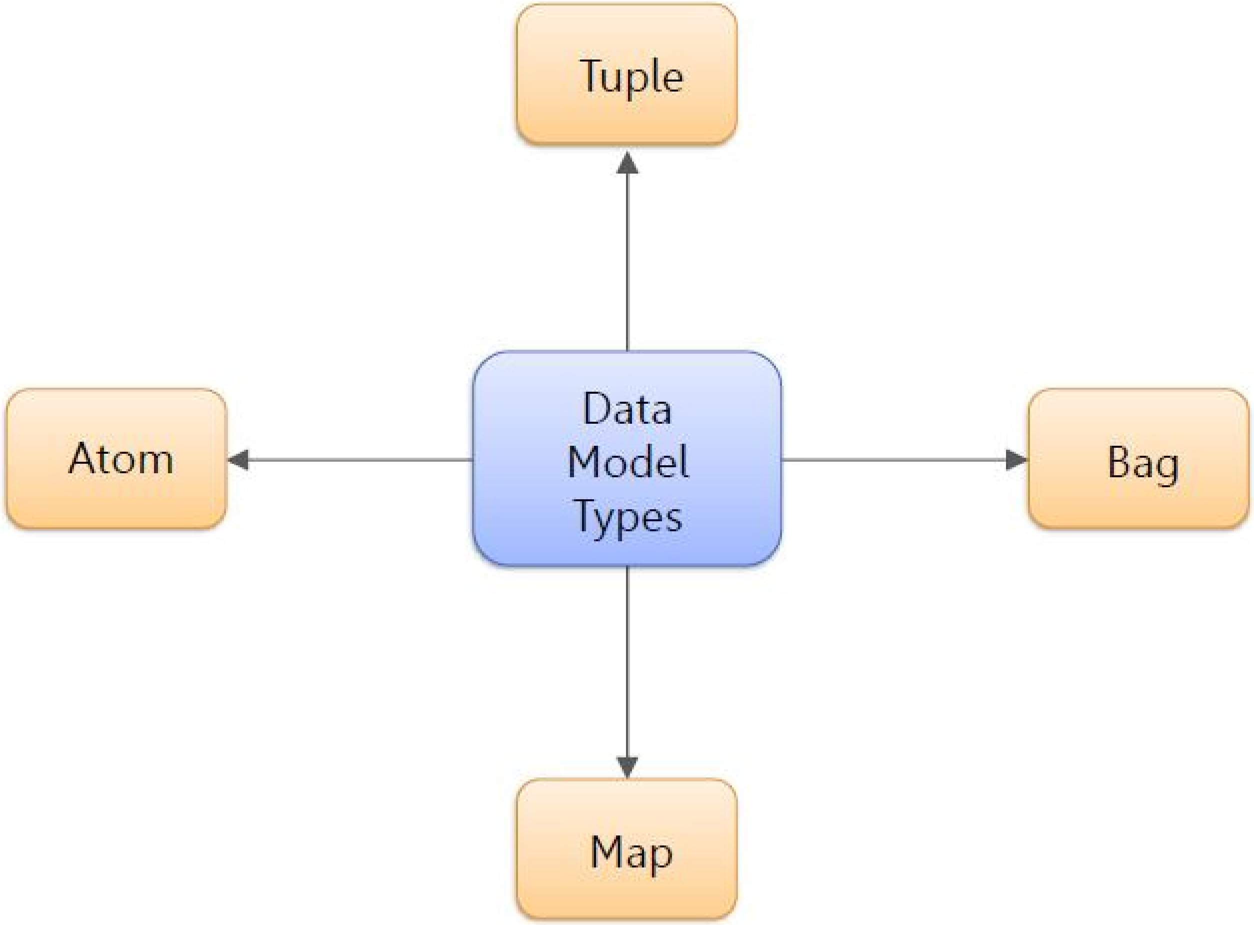


Pig Data Types

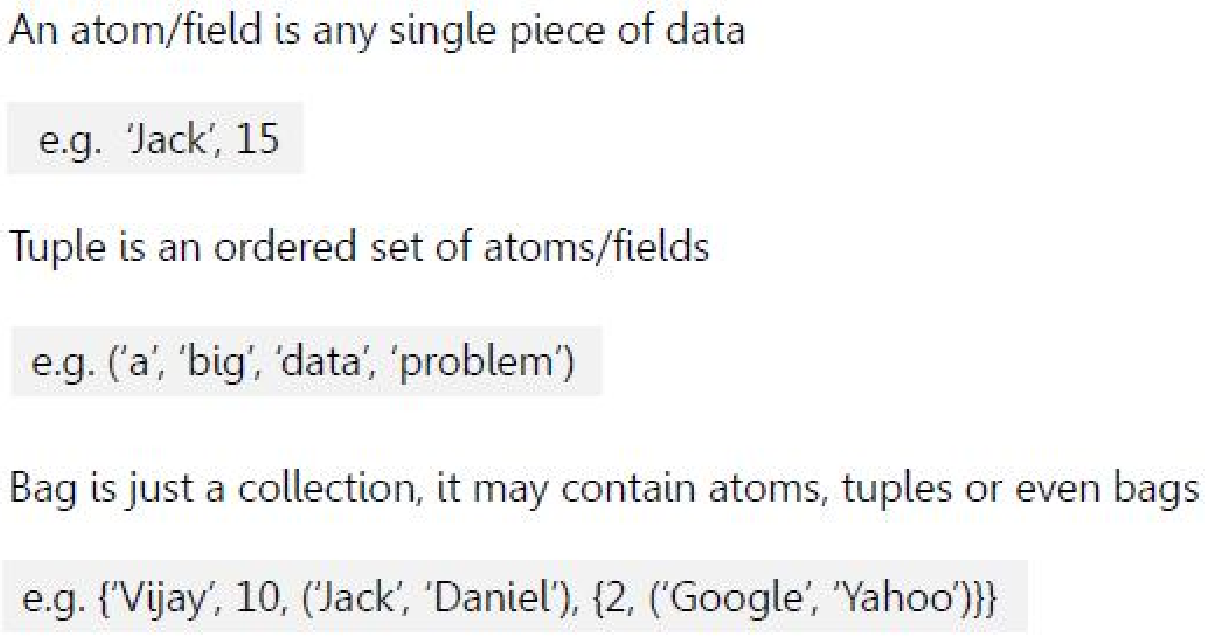


Pig Data Type	Implementing Class
Bag	org.apache.pig.data.DataBag
Tuple	org.apache.pig.data.Tuple
Map	java.util.Map<Object, Object>
Integer	java.lang.Integer
Long	java.lang.Long
Float	java.lang.Float
Double	java.lang.Double
Chararray	java.lang.String
Bytearray	byte[]

Complex Data Types



Complex Data Types



An atom/field is any single piece of data

e.g. ‘Jack’, 15

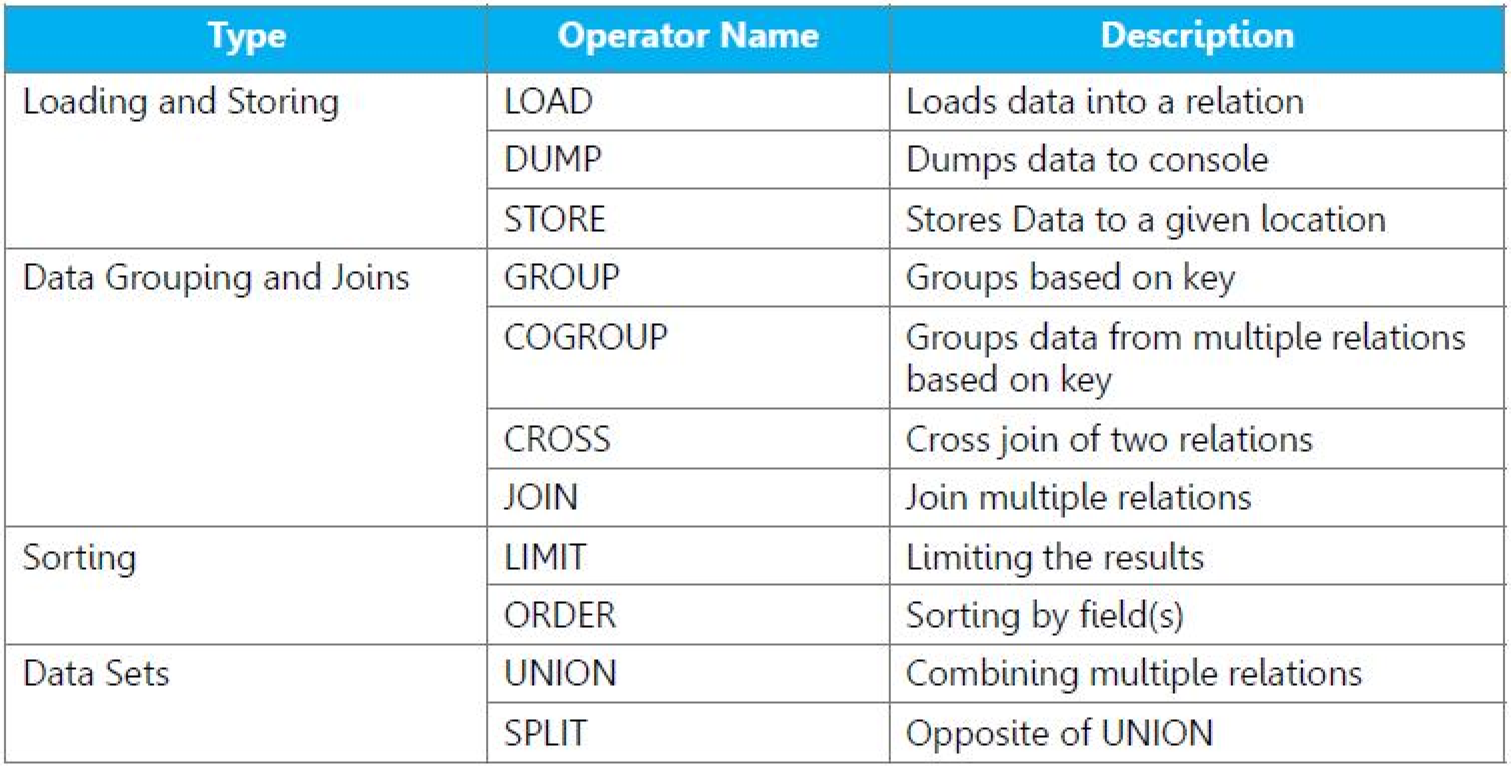
Tuple is an ordered set of atoms/fields

e.g. (‘a’, ‘big’, ‘data’, ‘problem’)

Bag is just a collection, it may contain atoms, tuples or even bags

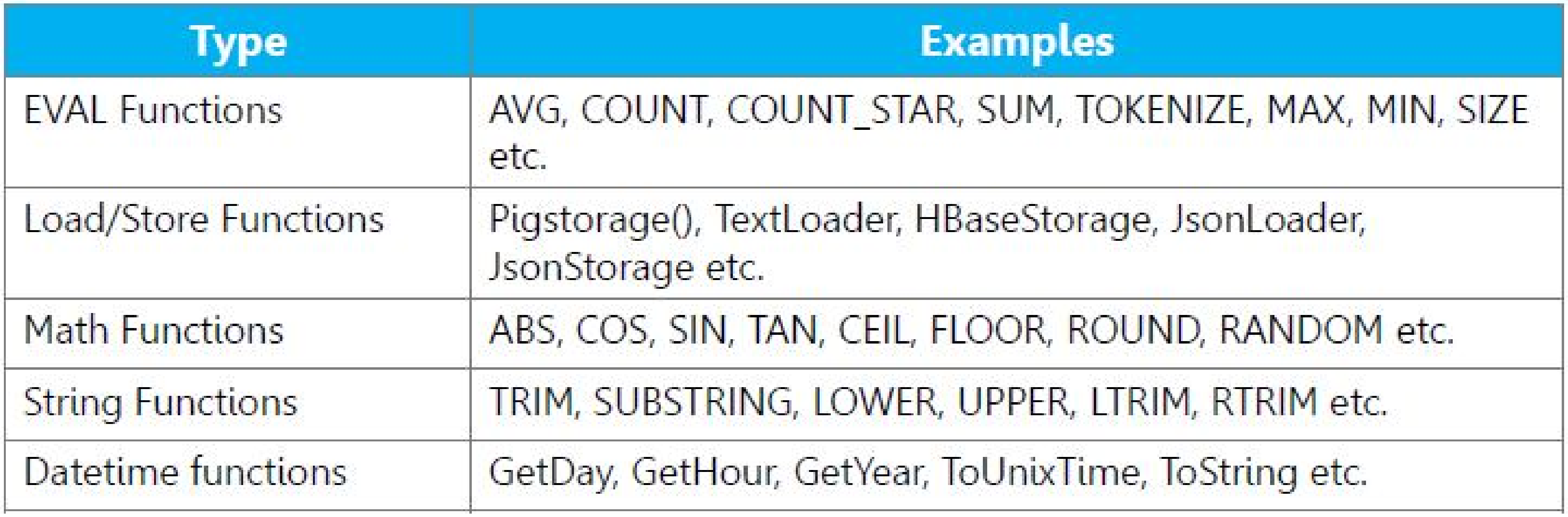
e.g. {‘Vijay’, 10, (‘Jack’, ‘Daniel’), {2, (‘Google’, ’Yahoo’)}}

Pig Operators



Type	Operator Name	Description
Loading and Storing	LOAD	Loads data into a relation
	DUMP	Dumps data to console
	STORE	Stores Data to a given location
Data Grouping and Joins	GROUP	Groups based on key
	COGROUP	Groups data from multiple relations based on key
	CROSS	Cross join of two relations
	JOIN	Join multiple relations
Sorting	LIMIT	Limiting the results
	ORDER	Sorting by field(s)
Data Sets	UNION	Combining multiple relations
	SPLIT	Opposite of UNION

Pig built-in Functions



Type	Examples
EVAL Functions	AVG, COUNT, COUNT_STAR, SUM, TOKENIZE, MAX, MIN, SIZE etc.
Load/Store Functions	Pigstorage(), TextLoader, HBaseStorage, JsonLoader, JsonStorage etc.
Math Functions	ABS, COS, SIN, TAN, CEIL, FLOOR, ROUND, RANDOM etc.
String Functions	TRIM, SUBSTRING, LOWER, UPPER, LTRIM, RTRIM etc.
Datetime functions	GetDay, GetHour, GetYear, ToUnixTime, ToString etc.

Thank You!