Introduction to Big Data

ECAP456

Dr. Rajni Bhalla
Associate Professor

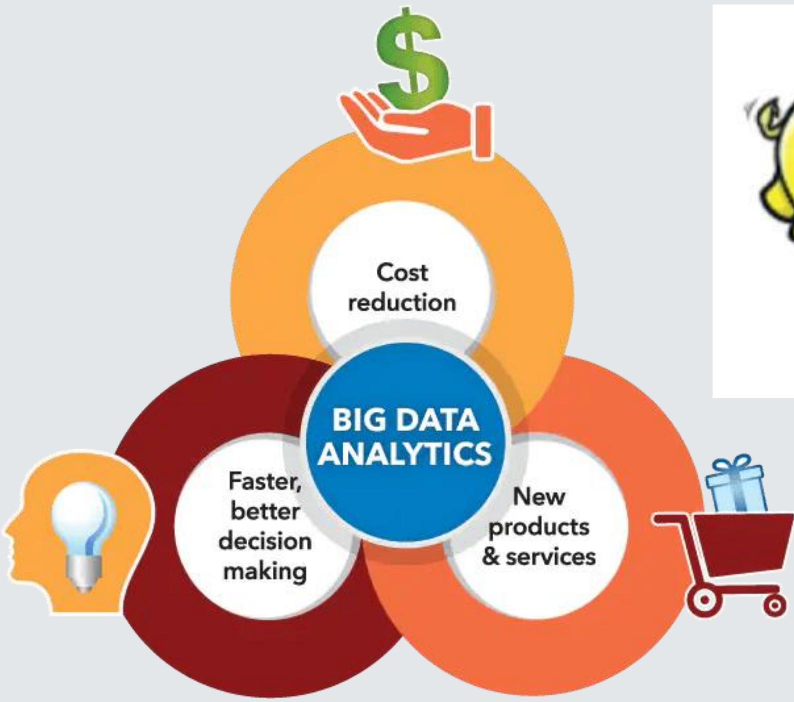
Learning Outcomes

After this lecture, you will be able to

- explore concepts of Apache pig
- learn Architecture of Apache Pig
- understand Pig-Latin data types
- learn Application of Pig
- explore features of pig

Introduction

Introduction



Introduction

- Challenging to write and
- Maintain these lengthy Java codes

Introduction

Apache Pig

Introduction

Apache Pig

Developed By

Introduction

For people who don't know how to programme in Java, this is a godsend.

Apache Pig

Introduction

For people who don't know how to programme in Java, this is a godsend.

Most Popular

Apache Pig

Introduction

For people who don't know how to programme in Java, this is a godsend.

Apache Pig

Most Popular

flexibility,

Introduction

For people who don't know how to programme in Java, this is a godsend.

Most Popular

Apache Pig

flexibility,

reduces code complexity

Introduction

For people who don't know how to programme in Java, this is a godsend.

Most Popular

Apache Pig

flexibility,

reduces code complexity

requires less effort

Map Reduce vs. Apache Pig

Why the name PIG?

Map Reduce vs. Apache Pig

Why the name PIG?

Architecture of PIG

Components

Architecture of PIG

Components

Pig Latin data model

Architecture of PIG

Components

Pig Latin data model

Pig Job Execution Flow in depth

Architecture of PIG

Pig Latin

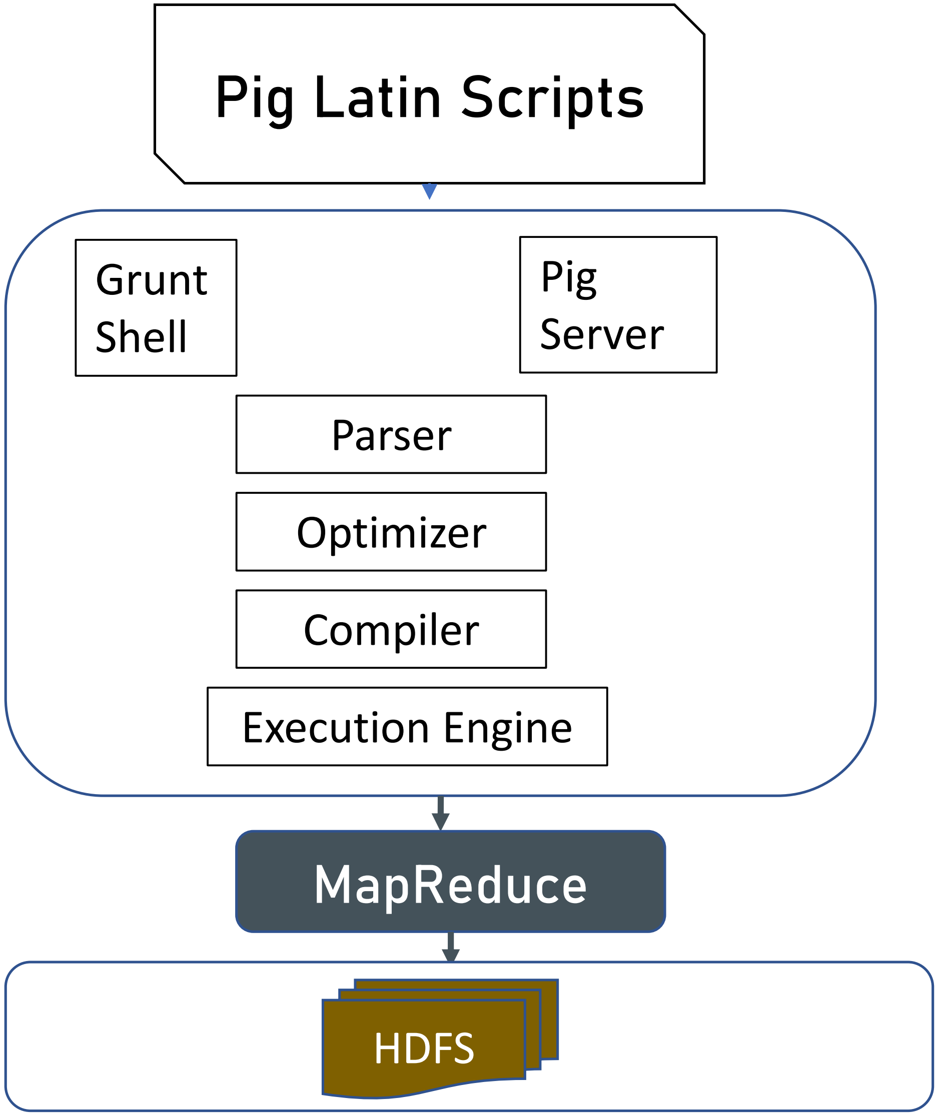
```
asianCountrys = filter countrys by region matches 'Asia';
joined = join customers by country_id, asianCountrys by country_id;
grouped = group joined by country_name;
agged = foreach grouped generate group, COUNT(joined.customers::cust_id);
morethan500cust = filter agged by $1 > 500;
ordered order morethan500cust by $1 desc;

dump ordered;
```

SQL or Hive QL

```
SELECT country_name,COUNT(cust_id) AS cust_count
  FROM countries co
  JOIN customers cu
    ON (co.country_id=cu.country_id)
 WHERE country_region='Asia'
 GROUP BY country_name
HAVING COUNT(cust_id)>500
 ORDER BY cust_count DESC
```

Architecture of PIG



Apache Pig Components

Parser

Category	Command	Description
Hadoop Filesystem	cat	Prints the contents of one or more files
	cd	Changes the current directory
	copyFromLocal	Copies a local file or directory to a Hadoop filesystem
	copyToLocal	Copies a file or directory on a Hadoop filesystem to the local filesystem
	cp	Copies a file or directory to another directory
	fs	Accesses Hadoop's filesystem shell
	ls	Lists files
	mkdir	Creates a new directory
	mv	Moves a file or directory to another directory
	pwd	Prints the path of the current working directory
	rm	Deletes a file or directory
	rmf	Forcibly deletes a file or directory (does not fail if the file or directory does not exist)
Hadoop MapReduce	kill	Kills a MapReduce job
Utility	exec	Runs a script in a new Grunt shell in batch mode
	help	Shows the available commands and options
	quit	Exits the interpreter
	run	Runs a script within the existing Grunt shell
	set	Sets Pig options

Commands

Apache Pig Components

Syntax check

Parser

Apache Pig Components

Syntax check

Type check

Parser

Apache Pig Components

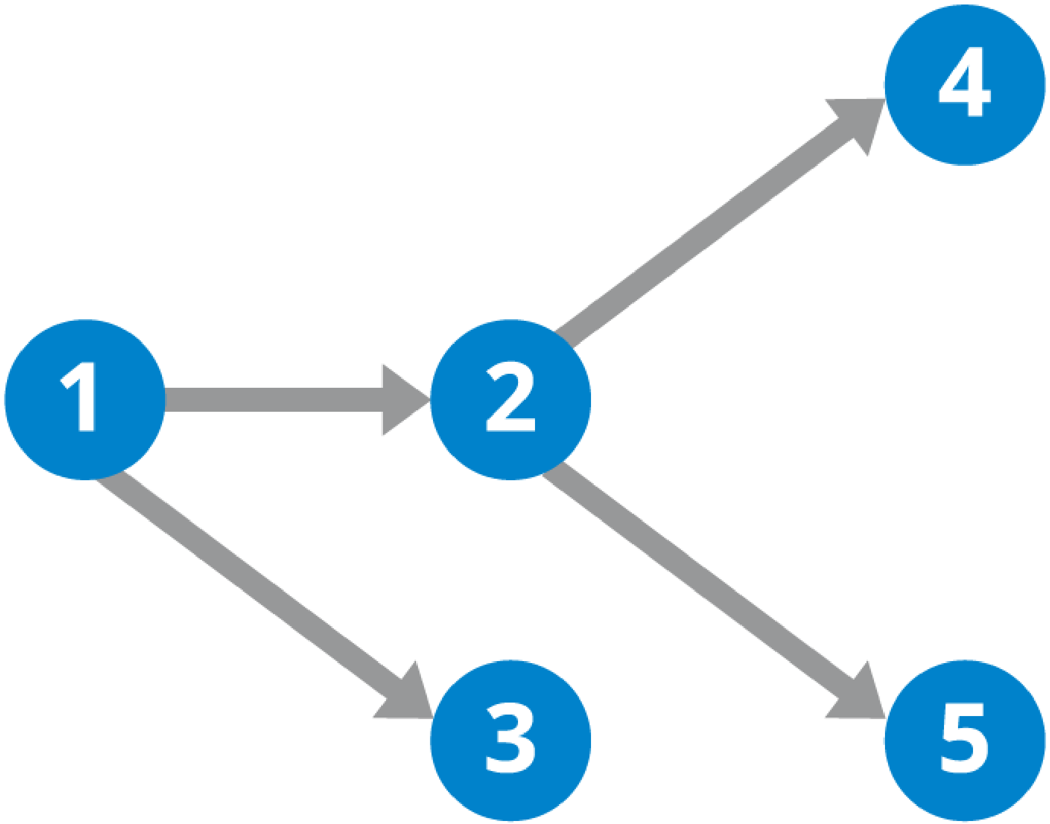
Syntax check

Type check

Parser

DAG(Directed Acyclic Graph)

Apache Pig Components



DAG(Directed Acyclic Graph)

Apache Pig Components

Optimizer

Projection

Push Down

Apache Pig Components

Compiler

Apache Pig Components

Compiler

Compiles

Apache Pig Components

Compiler

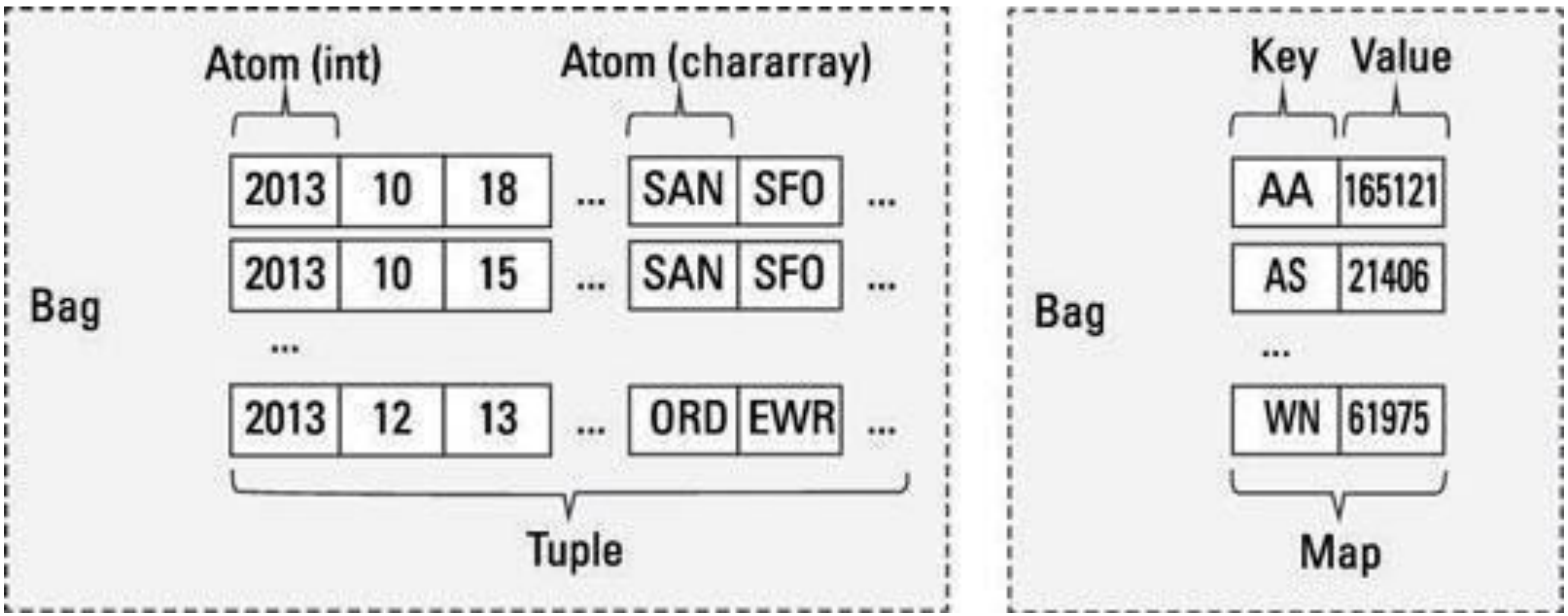
Compiles

Optimized logical plan

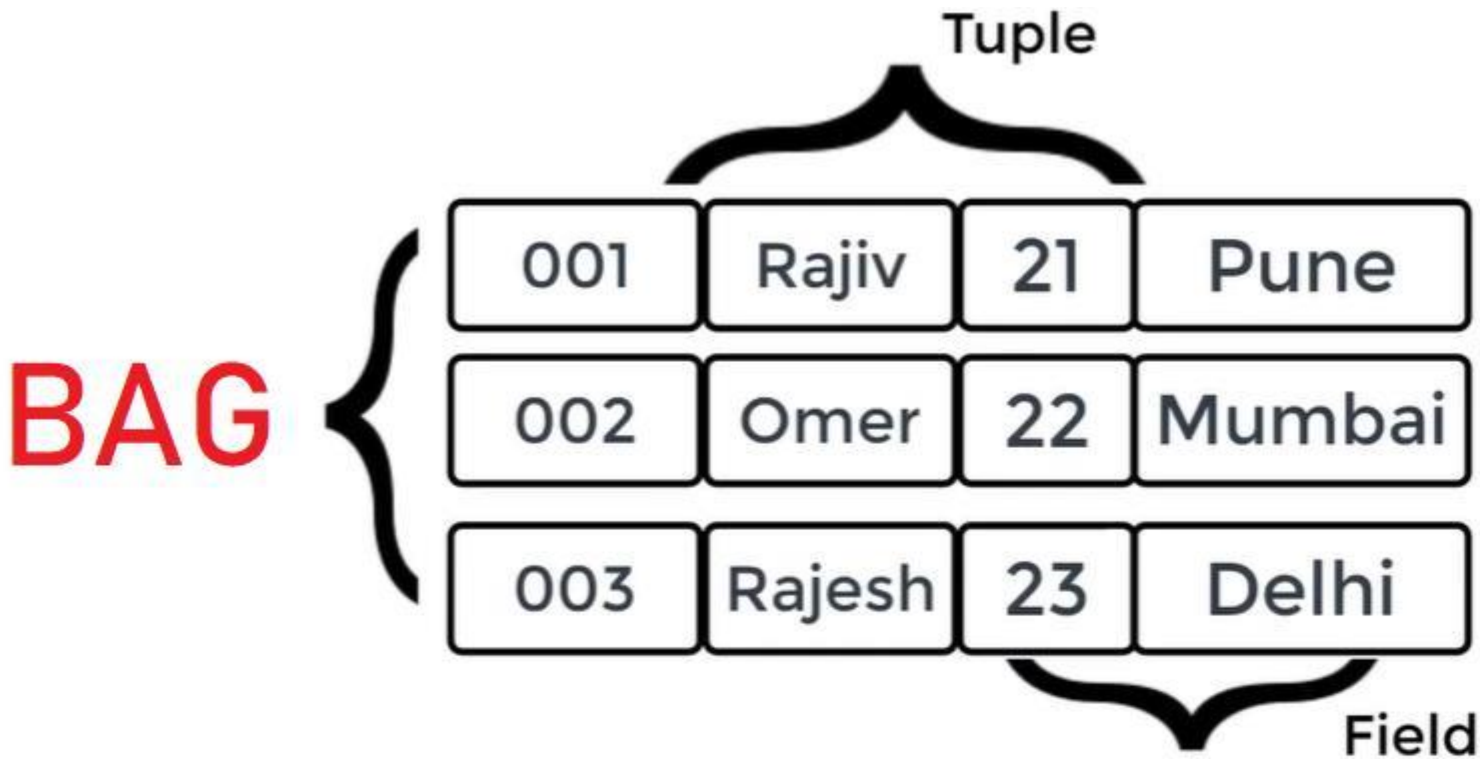
Apache Pig Components

Execution
engine

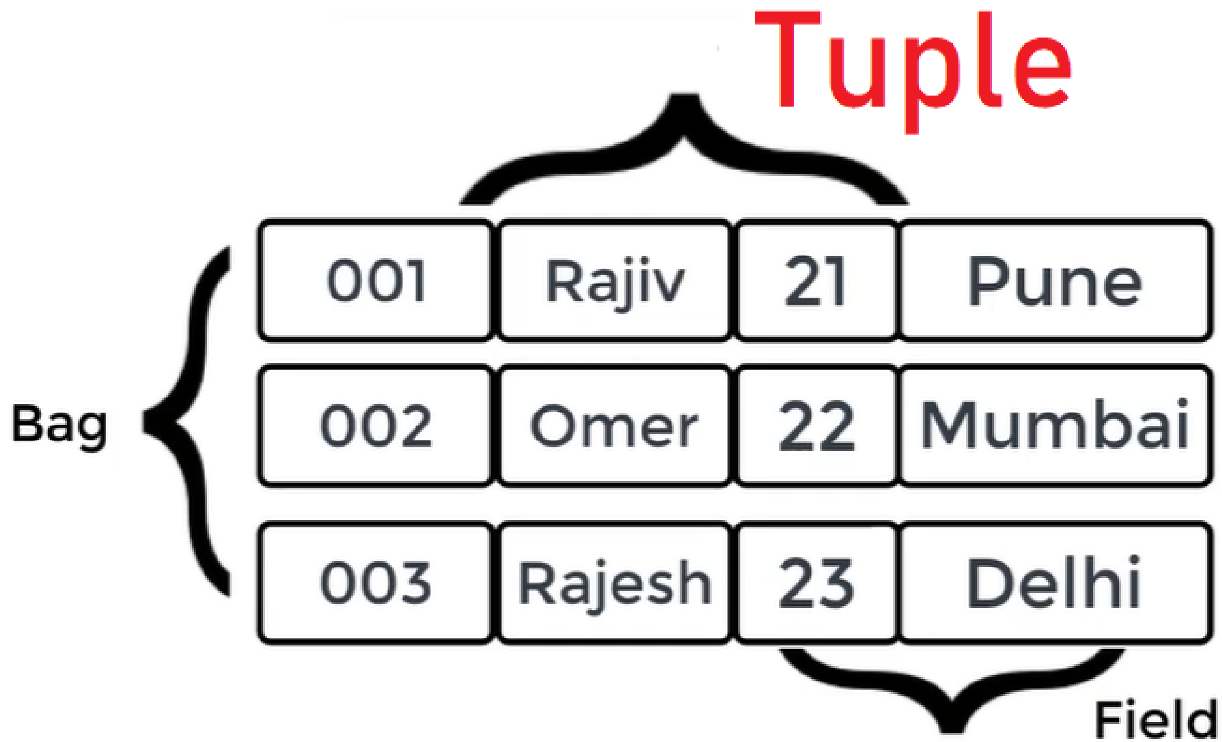
Pig Latin Data Model



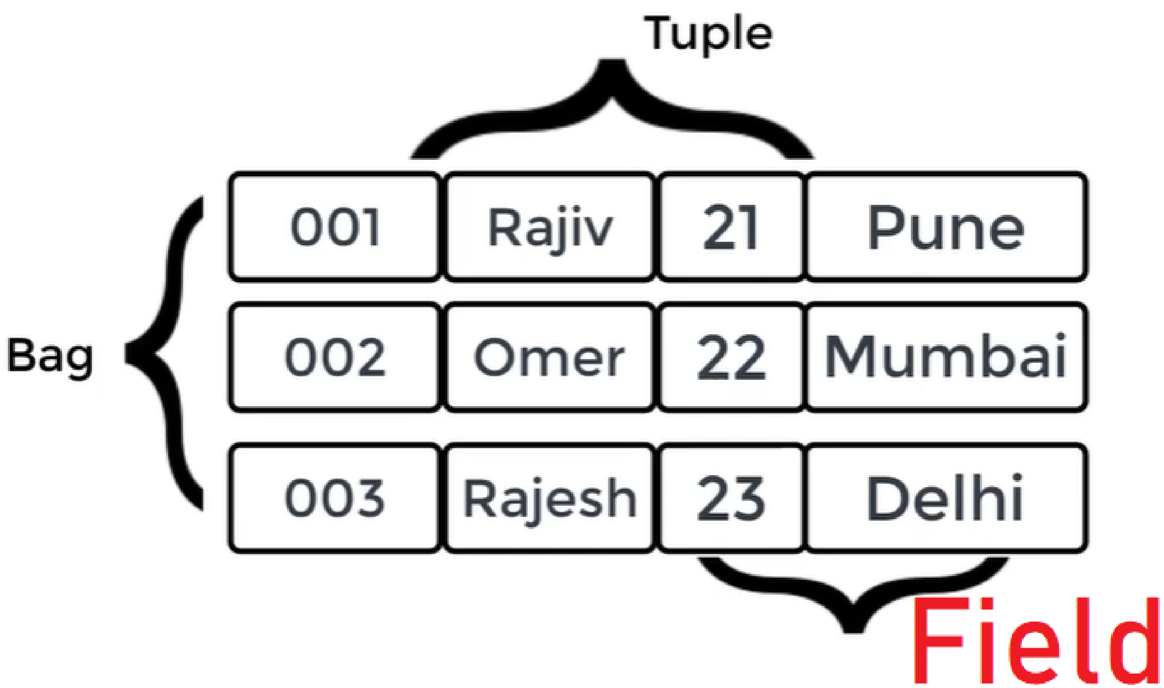
Pig Latin Data Model



Pig Latin Data Model



Pig Latin Data Model



Pig Latin Data Type

Data Type	Description	Example
Int	Represent a signed 32 bit integer	8
long	Represent a signed 64 bit integer	5L
float	Represent a signed 32 bit integer	5.5L
double	Represent a signed 64 bit integer	10.5L

Pig Latin Data Type

Data Type	Description	Example
Chararray	Represents a character array(string) in Unicode UTF-8 format	‘Online Classes’
Bytearray	Represents a Byte array(blob)	
Boolean	Represents a Boolean values	True/false

Pig Latin Data Type

Data Type	Description	Example
Datetime	Represents a Date-time.	1970-01-01T00:00:00.000+00:00
BigInteger	Represents a java Biginteger	60708090709
Bigdecimal	Represents a Bigdecimal	185.98376256272893883

Pig Latin Data Type

Complex Types

Data Type	Description	Example
Tuple	A tuple is an ordered set of fields	(raja,30)
Bag	A bag is a collection of tuples	{(raju,30),(Mohammad,45)}
Map	A Map is a set of key-value pairs	[‘name’#’Raju’,’age’#30]

Pig Latin – Arithmetic Operators

Operator
+ **Addition**
- **Subtraction**
* **Multiplication**
/ **Division**
% **Modulus**
Bincond
Case

Pig Latin – Arithmetic Operators

Bincond:

- Evaluates the Boolean operators. It has three operands as shown below.
- variable x = (expression) ? value1 if true : value2 if false.

Pig Latin – Arithmetic Operators

Bincond:

- b = (a == 1)? 20: 30;

if a = 1 the value of b is 20.

if a!=1 the value of b is 30.

Pig Latin – Arithmetic Operators

Case – The case operator is equivalent to nested bincond operator.

```
CASE f2 % 2
WHEN 0 THEN 'even'
WHEN 1 THEN 'odd'
END
```

Pig Latin – Comparison Operators

==Equal
!=Not Equal
>Greater than
<Less than
>=Greater than or equal to
<=Less than or equal to
Matches(Pattern matching)

Pig Latin – Type Construction Operators

Operator- Description
() – Tuple constructor operator
{} - Bag constructor operator
[] - Map constructor operator

Pig Latin – Relational Operators

Operator- Description
Loading and Storing
LOAD
STORE
Filtering
FILTER
DISTINCT
FOREACH, GENERATE
STREAM
Grouping and Joining
JOIN
COGROUP
GROUP
CROSS

Pig Latin – Relational Operators

Sorting
ORDER
LIMIT
Combining and Splitting
UNION
SPLIT
Diagnostic Operators
DUMP
DESCRIBE
EXPLAIN
ILLUSTRATE

Applications of Apache Pig:

- For exploring large datasets Pig Scripting is used.
- Provides the supports across large data-sets for Ad-hoc queries.
- In the prototyping of large data-sets processing algorithms.

Applications of Apache Pig:

- Required to process the time sensitive data loads.
- For collecting large amounts of datasets in form of search logs and web crawls.
- Used where the analytical insights are needed using the sampling.

Features of Apache Pig

- Rich set of operators
- Ease of programming
- Optimization opportunities
- Extensibility
- UDF's
- Handles all kinds of data

That's all for now...