Introduction to Big Data

ECAP456

Dr. Rajni Bhalla

Associate Professor

Learning Outcomes

After this lecture, you will be able to

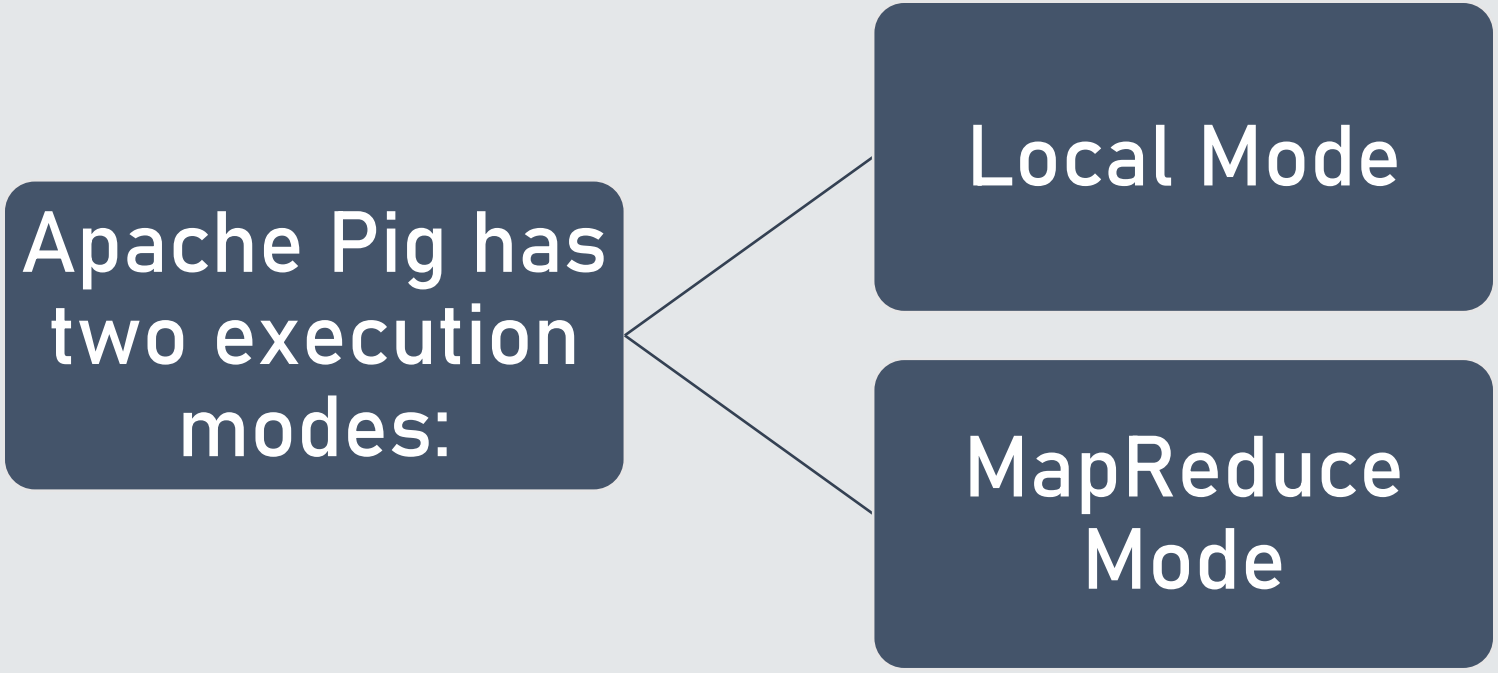
- Operators in Apache Pig

Introduction

- It is a high-level procedural language for querying large data sets using Hadoop and the Map Reduce Platform.
- It is a Java package, where the scripts can be executed from any language implementation running on the JVM.
- It simplifies the use of Hadoop by allowing SQL-like queries to a distributed dataset
- It backs many relational features like Join, Group and Aggregate.
- it does have many features common with ETL tools.

What is Apache Pig Latin?

Apache Pig Execution Modes

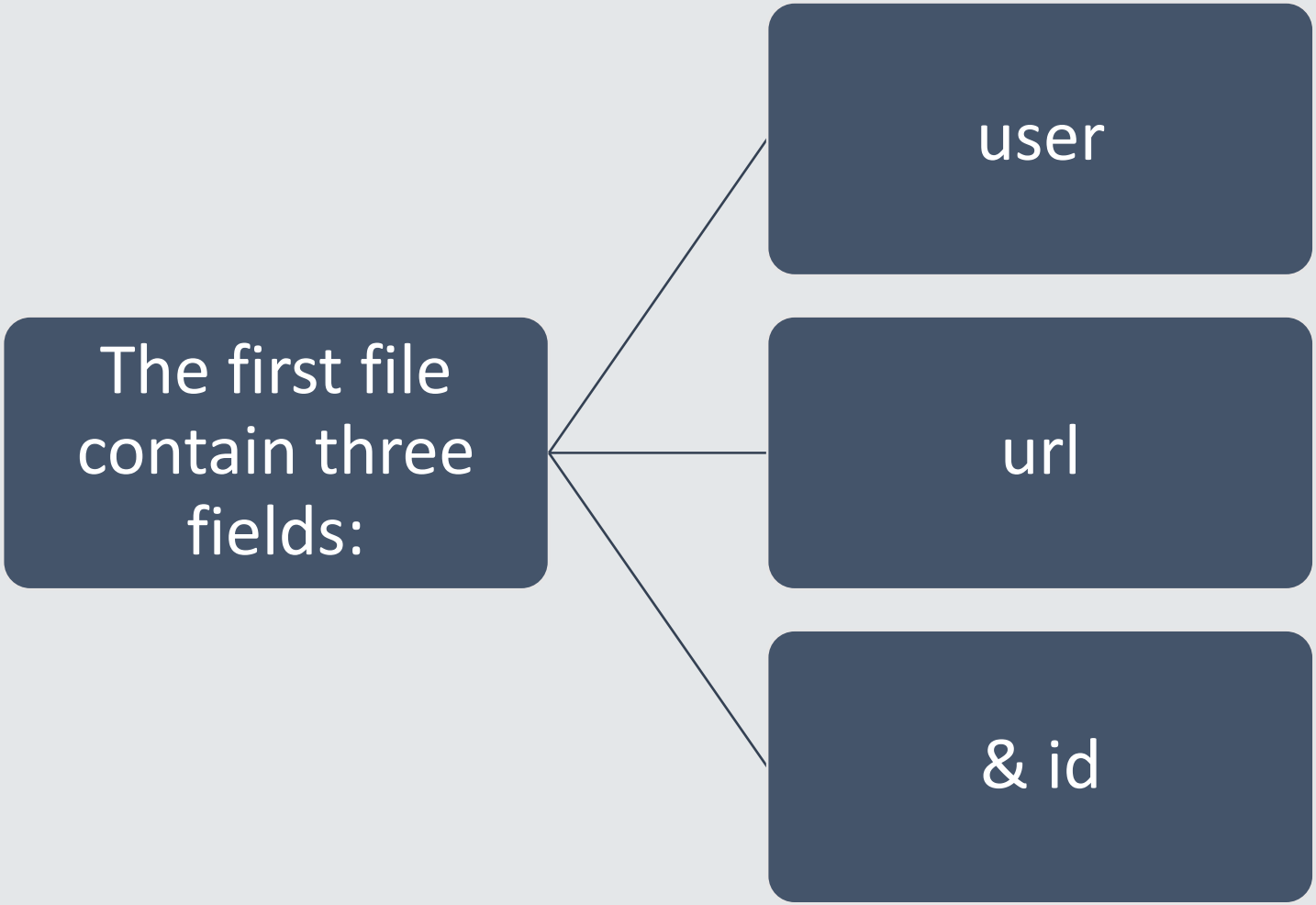


Local Mode

MapReduce Mode

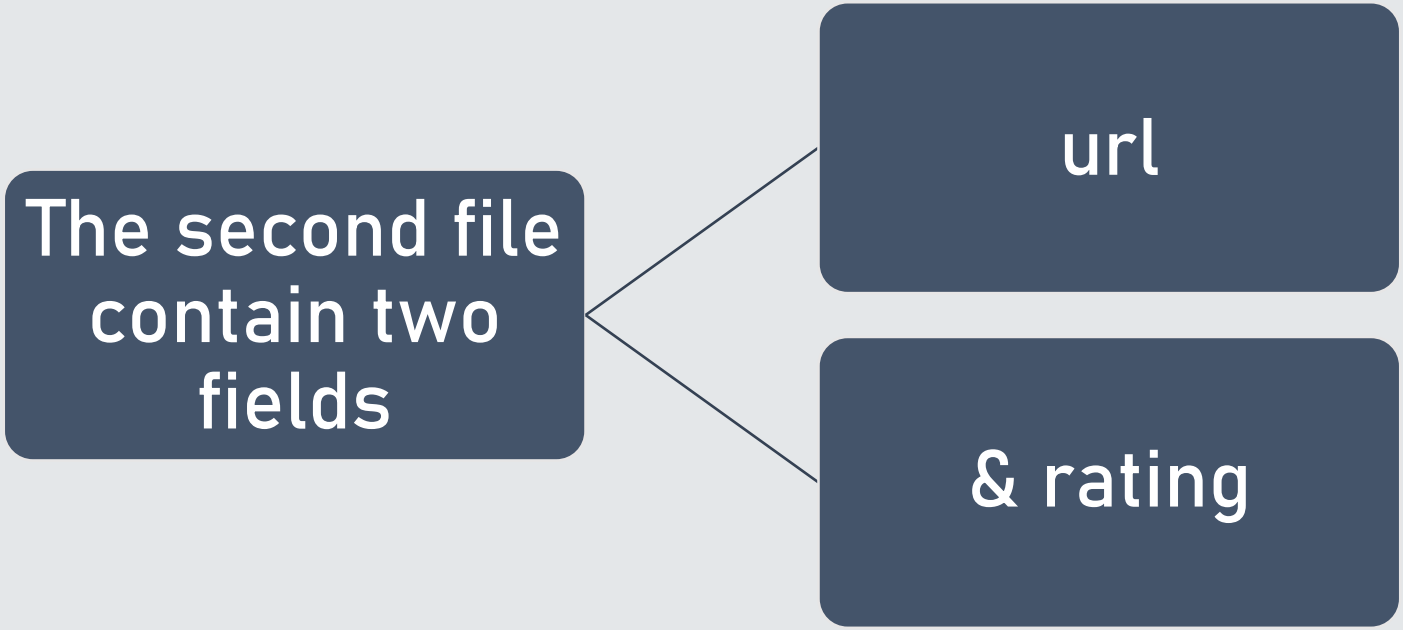
Apache Pig Operators

Apache Pig Operators

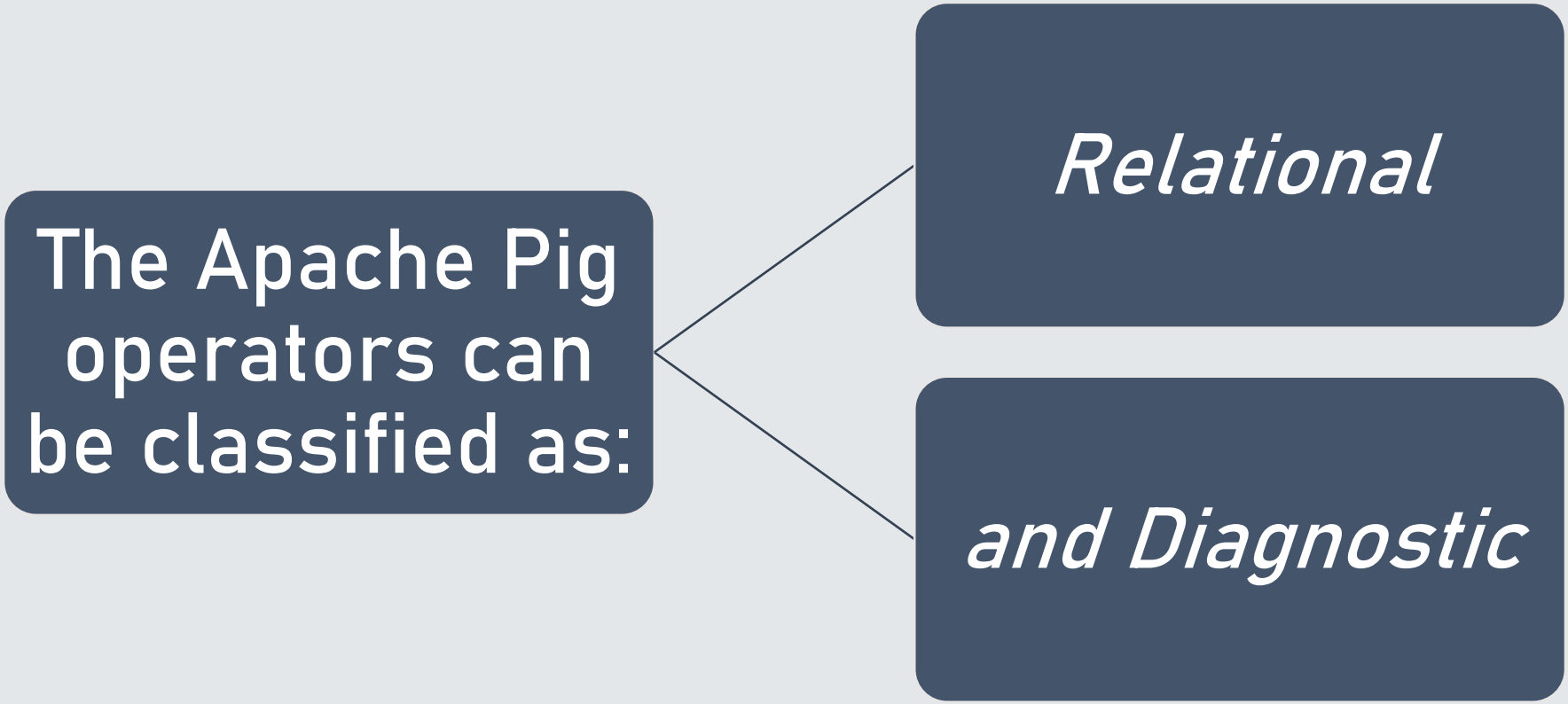


Apache Pig Operators

- .



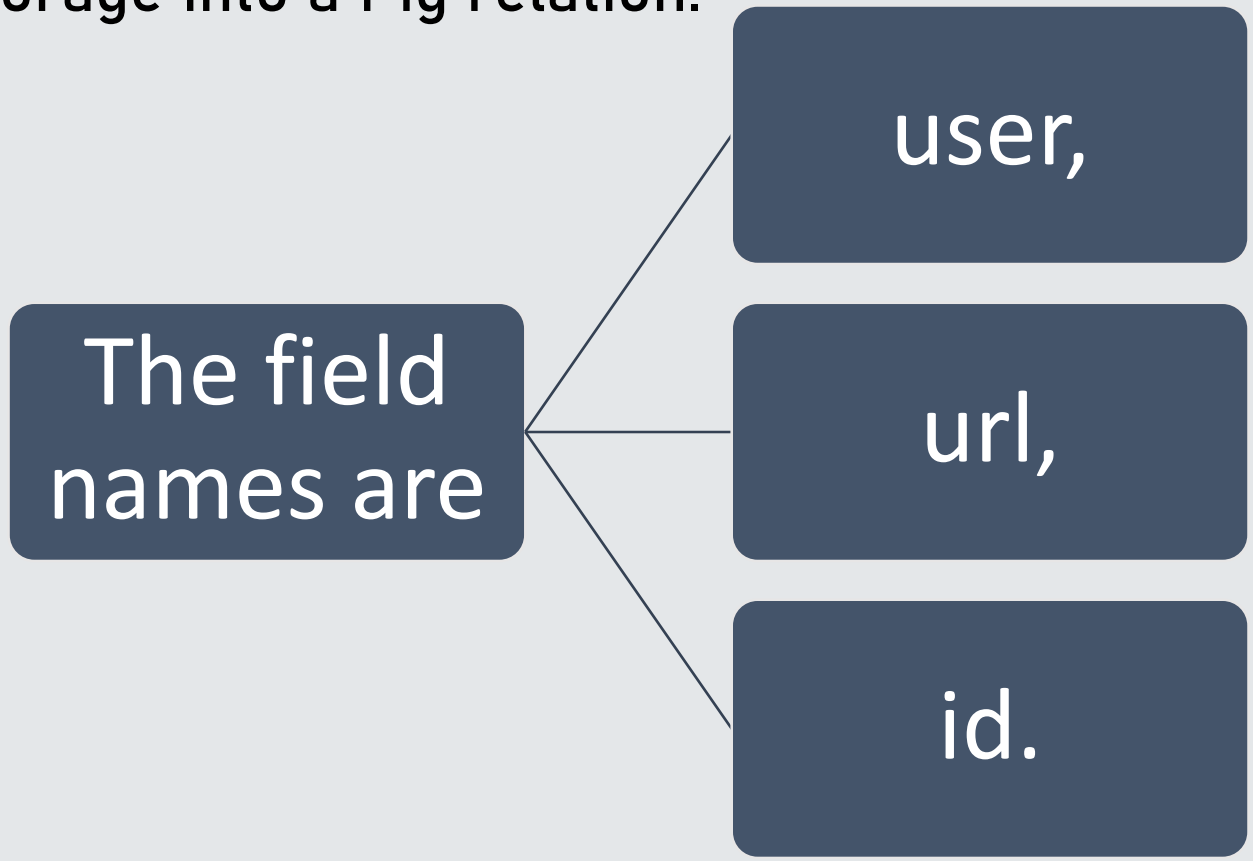
Classification of Apache Pig operators



Relational Operators:

Load Operators:

LOAD operator is used to load data from the file system or HDFS storage into a Pig relation.



Load Operators:

Foreach

Foreach

Filter

Filter

Join

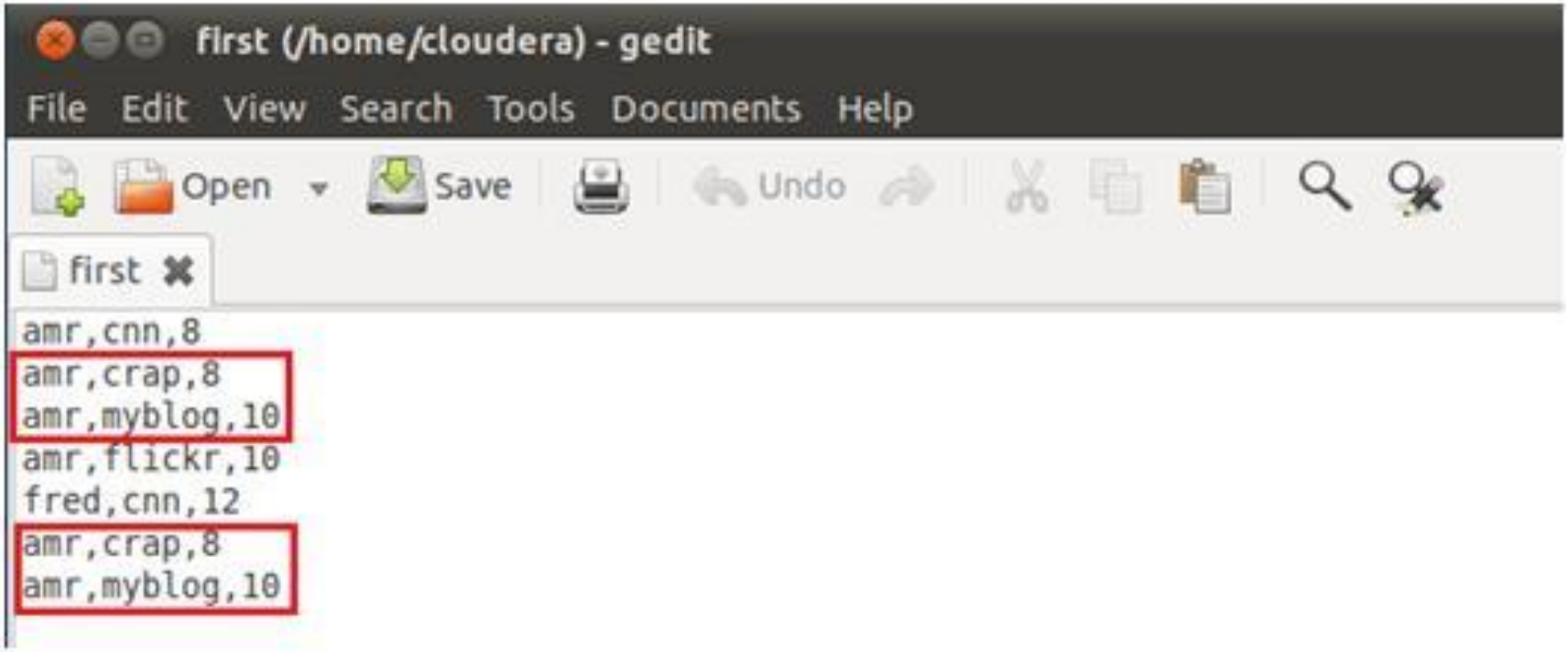
- JOIN operator is used to perform an inner, equijoin join of two or more relations based on common field values.
- The JOIN operator always performs an inner join.
- Inner joins ignore null keys, so it makes sense to filter them out before the join.

Order By

Order By

Distinct

Distinct


first (/home/cloudera) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
first
amr,cnn,8
amr,crap,8
amr,myblog,10
amr,flickr,10
fred,cnn,12
amr,crap,8
amr,myblog,10

Store

Store


That's all for now...