Apache Pig

What is Pig?

- Pig is a platform that creates Map-Reduce programs.
- Pig Latin is a language for this platform to express data processing.
  - Pig Latin is procedural
  - Each Step specifies single High-level data transformation.
  - Supports nested Data Model (Tuple in Tuple or Bag in Tuple etc.)
  - Extensible via use of User Defined Functions
- Can be used effectively for ad-hoc data analysis

Pig Timelines

- Pig started as a research project within Yahoo! in the summer of 2006.
- Joined Apache Incubator in September of 2007
- Our CDH distribution has Pig version 0.8.1
- Latest version of Pig is 0.11.1 (April 2013)

Running Pig

- Pig can be run in two modes
  - Local
    - pig –x local
  - Map-Reduce
    - Default mode
    - Needs following Environment variables set
      - HADOOP_CONF_DIR
      - PIG_CLASSPATH
      - Can control map-reduce properties using a Properties file passed as a command line argument when executing Pig Script
      - E.g.
        - » pig -P config/<<your properties file name>> somescript.pig
        - » <<your properties file name>> has a following parameter
          - *mapred.child.java.opts=-Xmx3072m*
- Pig has a interactive shell, Grunt, where we can enter Pig Latin command manually.

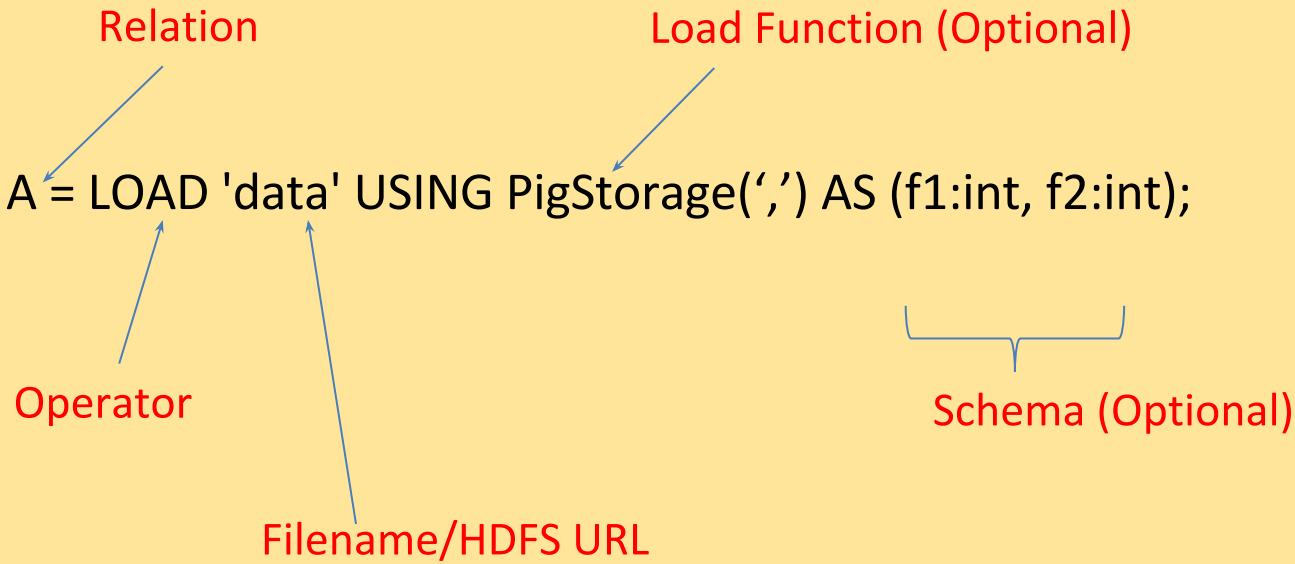
Pig Data Types

- Scalar
  - int, long, double, float, bytearray, chararray
- Complex
  - Tuple
    - Ordered Set of Fields
    - Represented as ()
  - Bag
    - Collection of Tuples
    - Represented as {}
  - Map
    - Set of Key value pairs where value can be either scalar or complex datatype.
    - Represented as [key#value]

Pig Latin Data Flow

- Load the data using Loader
- Process the data using Relational Operators
- Either Dump OR Store the result.
- *Note: Dump OR Store is necessary to initiate execution of Map-Reduce Job*.

Pig Latin Expression



Load/Store Function

- Load/Store determines how data flows into Pig and comes out of Pig.
- Pig provides Built-in load/store functions.
- PigStorage is the default load/store if “USING” is not used while loading the file.
- PigStorage use “TAB” as a default delimiter between fields.

Relational Operators

- FOREACH
  - To work with each tuple
- FILTER
  - Filter the Relation using values in Column
- DISTINCT
  - Remove Duplicate records
- COGROUP/GROUP
  - Collect the records using key from one or more inputs
- LIMIT
  - Limit number of Records
- JOIN
  - Join two or more relations based on the key

FLATTEN Operator

- Operator that changes structure of Tuples and Bags.
- Used in FOREACH
- Converts fields of tuple in place of tuple.
  - Input (a,(b,c))
  - Output (a,b,c)
- If Tuple has a Bag into it as a field then
  - Input – (a, {(b,c),(d,e)}
  - Output - (a,b,c) and (a,d,e)

Debugging

- DESCRIBE
  - Describes the Schema of a Relation
- ILLUSTRATE
  - Displays step by step execution of the statements
- DUMP
  - Dumps the output

User Defined Functions

- Way to extend the Pig Latin
- UDF extend EvalFunc<T>
- Implement "*abstract T exec(Tuple input)*"
- Register the JAR file that has the UDF using "REGISTER" in the script

References

- http://pig.apache.org/docs/r0.8.1/
- http://developer.yahoo.com/hadoop/tutorial/pigtutorial.html
- http://pig.apache.org/docs/r0.8.1/api/
- http://pig.apache.org/docs/r0.8.1/udf.html