Distributed Query Processing with Apache Pig

Prof. Pramod Bhatotia
https://dse.in.tum.de/bhatotia/

Limitations of MapReduce

- Graph algorithms
  - Pregel [SIGMOD '10], GraphX [OSDI'14]
- Iterative algorithms
  - Haloop [VLDB'10], CIEL [NSDI '11]
- Stream processing – Low latency
  - D-stream [SOSP'13], Naiad [SOSP'13], Storm, S4
- Low-level abstraction for common data analysis tasks!
  - Pig [SIGMOD'10], Shark [SIGMOD'13], DryadLINQ [OSDI'08]

Motivation for Pig

Programmers are lazy!

(they don't even wish to write Map and Reduce)

Data analysis tasks

- Common operations:
  - Filter, join, group-by, sort, etc.
- MapReduce offers a low-level primitive
  - Requires repeated re-implementation of these operators
- The power of abstraction!
  - Design once and reuse

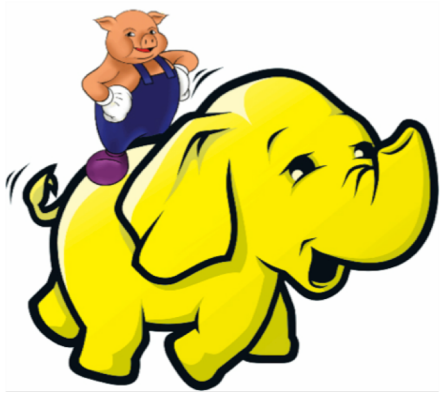
Pig Latin

Distributed dataflow queries

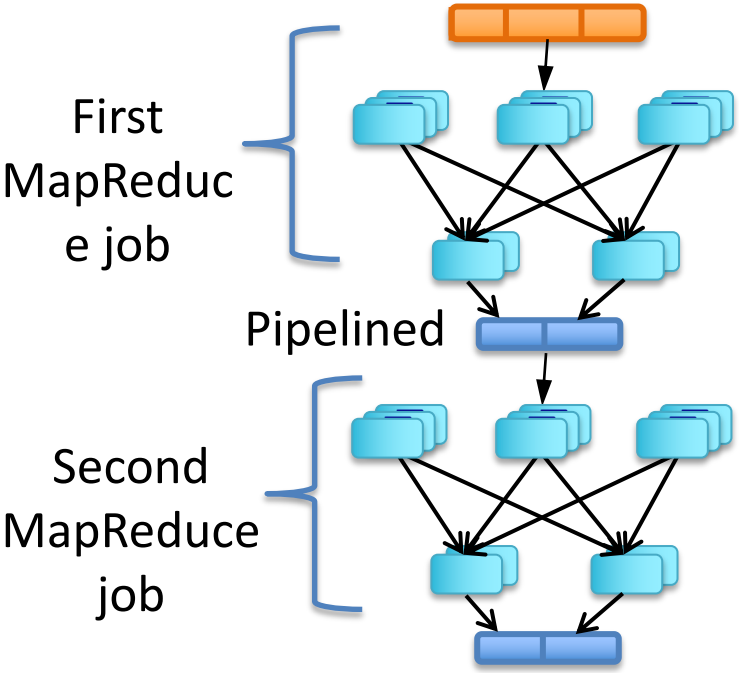
Pig Latin = SQL-kind queries + Distributed execution

Pig architecture

Pig Latin script

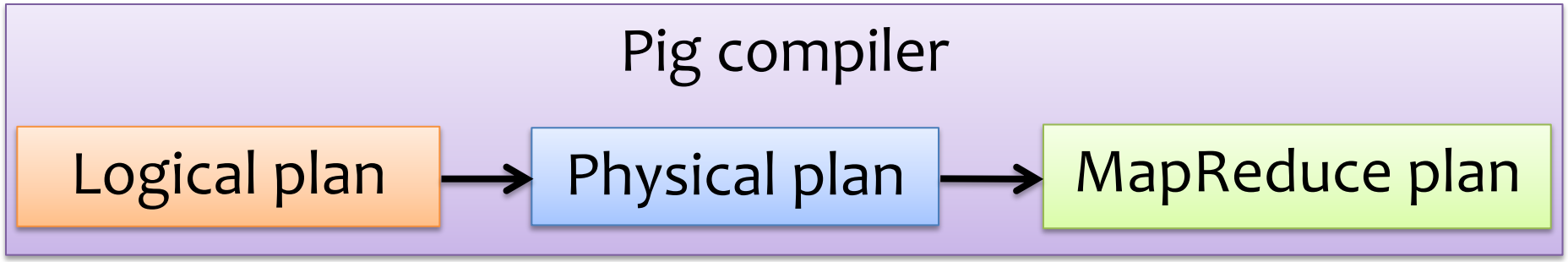


Pig compiler



MapReduce Runtime

Overview of the compilation process

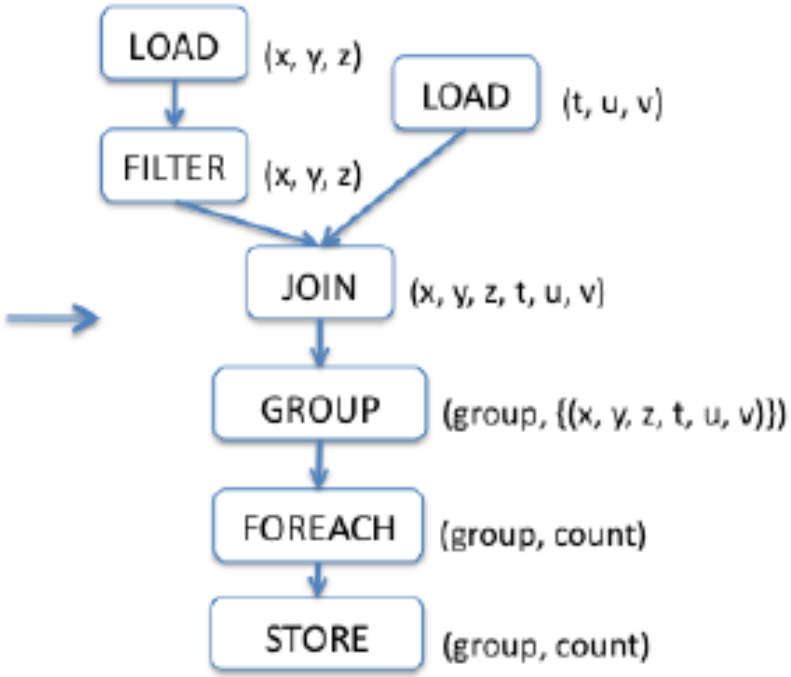


An example

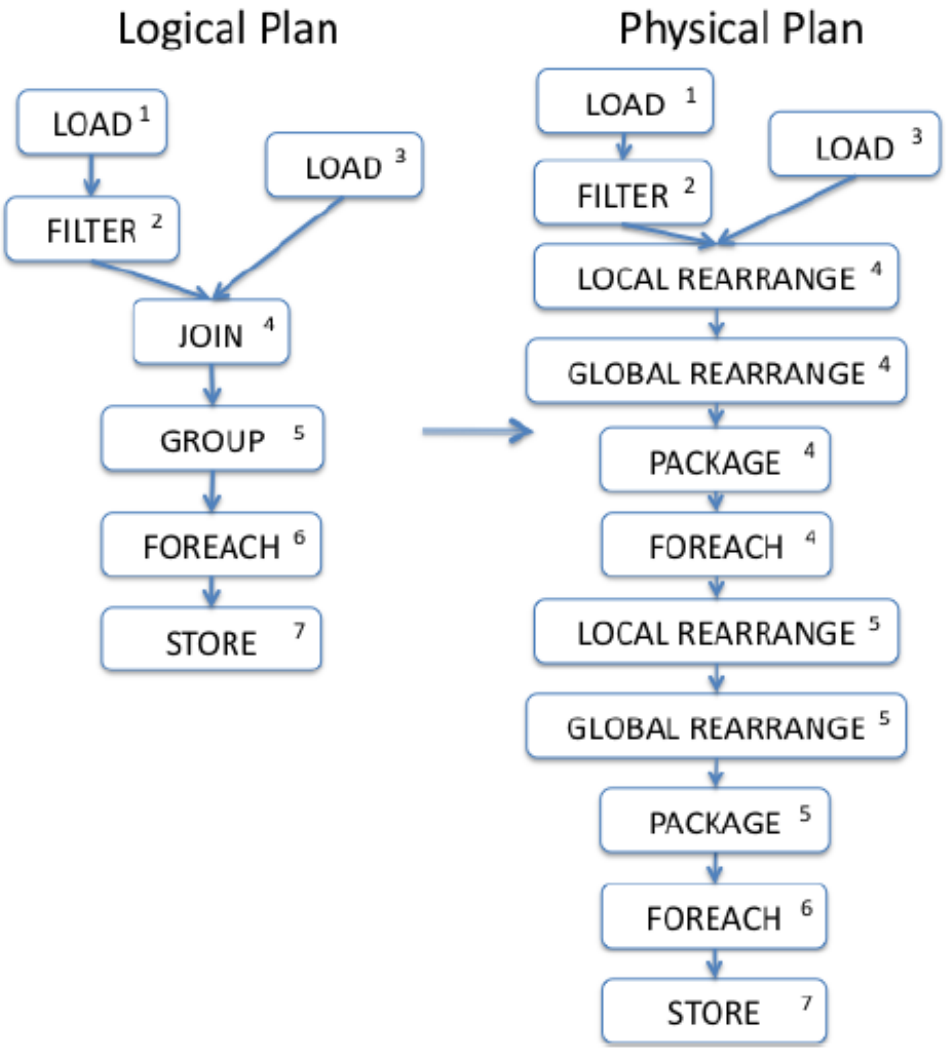
Pig Latin

```
A = LOAD 'file1' AS (x, y, z);
B = LOAD 'file2' AS (t, u, v);
C = FILTER A by y > 0;
D = JOIN C BY x, B BY u;
E = GROUP D BY z;
F = FOREACH E GENERATE
    group, COUNT(D);
STORE F INTO 'output';
```

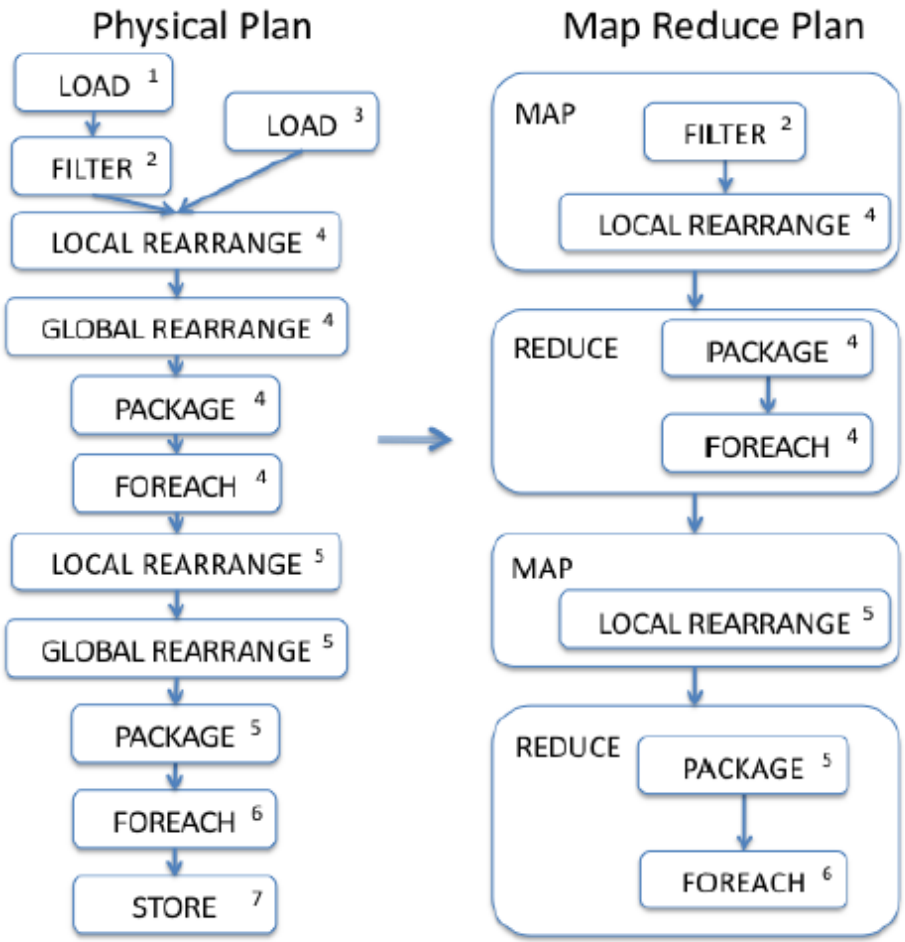
Logical Plan



Example: contd.



Example: contd.



Advantages of staged-compilation

- SQL query optimizations

- MapReduce specific optimizations

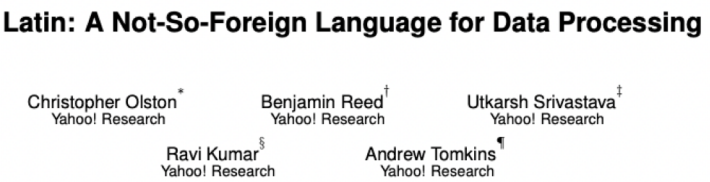
Refer Pig papers for details [SIGMOD '08, VLDB'09]

Related systems

- Apache HIVE
  - Built on top of MapReduce
- DryadLINQ [OSDI'08] or SCOPE [VLDB'08]
  - Built on top of Dryad
- Shark [SIGMOD'13]/ Spark SQL [SIGMOD'15]
  - Built on top of Spark

References

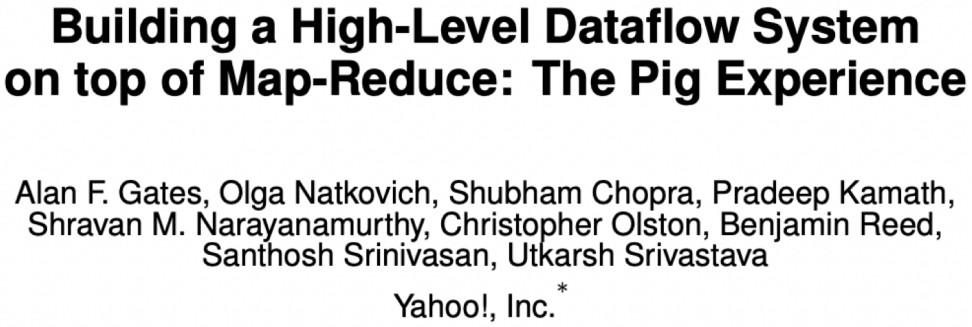
- Compulsory reading
  - Apache Pig [SIGMOD'08] http://infolab.stanford.edu/~olston/publications/sigmod08.pdf



Pig Latin: A Not-So-Foreign Language for Data Processing

Christopher Olston* Yahoo! Research
Benjamin Reed† Yahoo! Research
Utkarsh Srivastava‡ Yahoo! Research
Ravi Kumar§ Yahoo! Research
Andrew Tomkins¶ Yahoo! Research

  - Apache Pig [VLDB '09] http://infolab.stanford.edu/~olston/publications/vldb09.pdf



Building a High-Level Dataflow System on top of Map-Reduce: The Pig Experience

Alan F. Gates, Olga Natkovich, Shubham Chopra, Pradeep Kamath, Shravan M. Narayanamurthy, Christopher Olston, Benjamin Reed, Santhosh Srinivasan, Utkarsh Srivastava

Yahoo!, Inc.*

Summary

- Data-intensive computing with MapReduce
  - Data-parallel programming model
  - Runtime library to handle all low-level details
  - Pig: high-level abstraction for common tasks

- Resources:
  - Hadoop: http://hadoop.apache.org/
  - Spark: https://spark.apache.org/
  - Dryad: http://research.microsoft.com/en-us/projects/dryad/