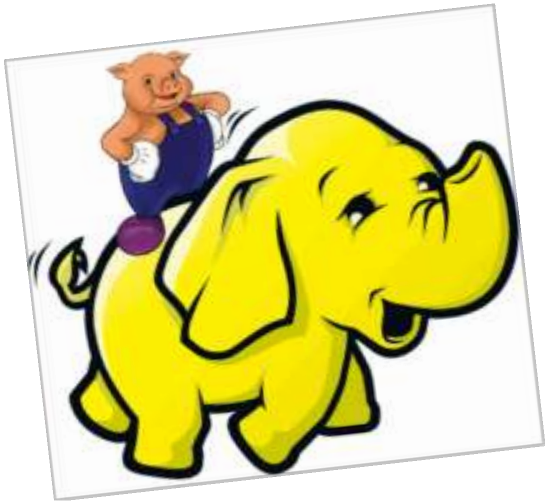
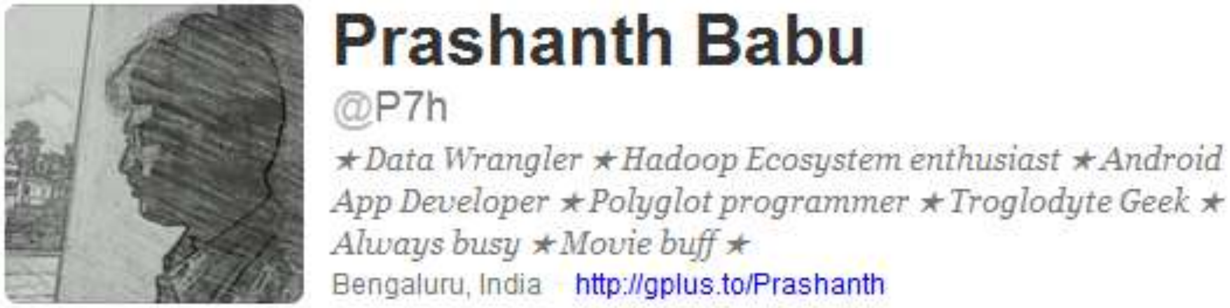


Introduction to Pig

Prashanth Babı

http://twitter.com/P7h

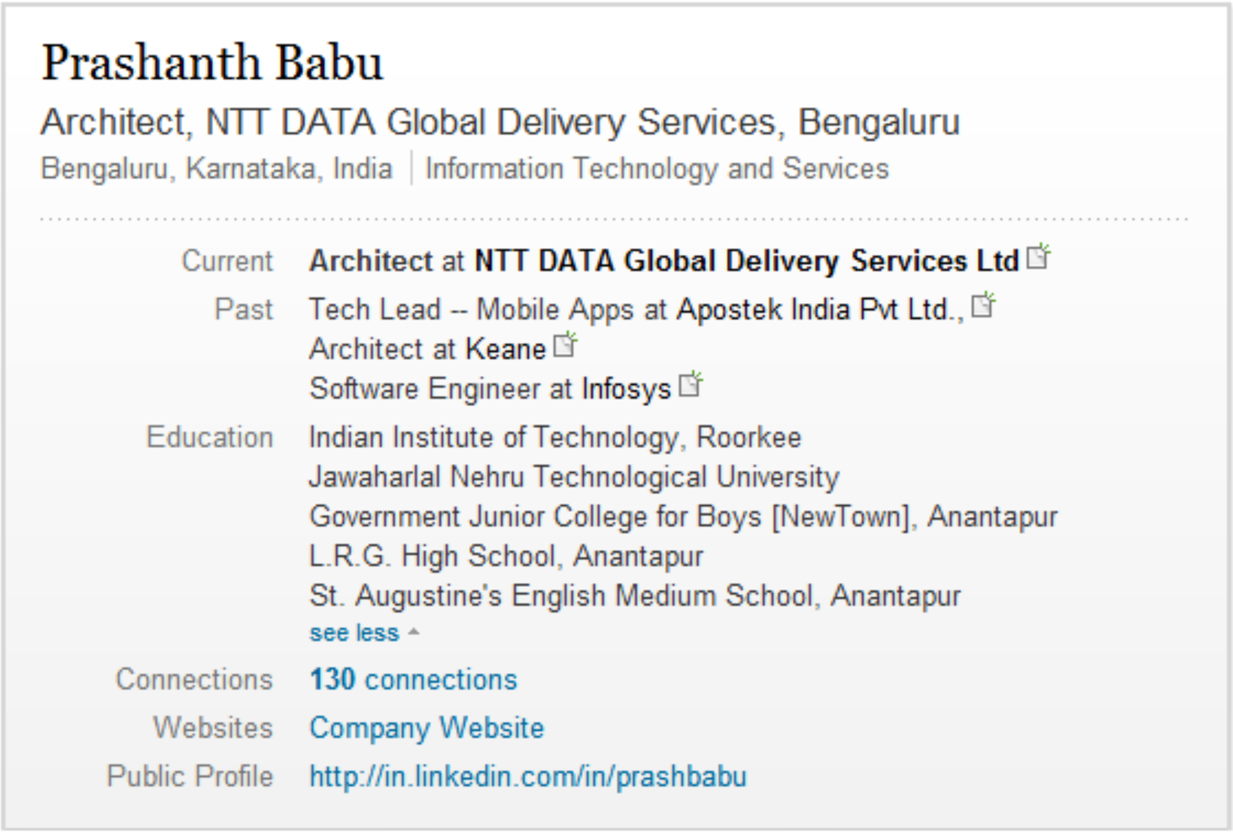


Prashanth Babu

@P7h

★ Data Wrangler ★ Hadoop Ecosystem enthusiast ★ Android App Developer ★ Polyglot programmer ★ Troglodyte Geek ★ Always busy ★ Movie buff ★

Bengaluru, India · http://gplus.to/Prashanth



Prashanth Babu

Architect, NTT DATA Global Delivery Services, Bengaluru

Bengaluru, Karnataka, India | Information Technology and Services

Current	**Architect at NTT DATA Global Delivery Services Ltd**
Past	Tech Lead -- Mobile Apps at Apostek India Pvt Ltd., Architect at Keane Software Engineer at Infosys
Education	Indian Institute of Technology, Roorkee Jawaharlal Nehru Technological University Government Junior College for Boys [NewTown], Anantapur L.R.G. High School, Anantapur St. Augustine's English Medium School, Anantapur see less ▲
Connections	130 connections
Websites	Company Website
Public Profile	http://in.linkedin.com/in/prashbabu

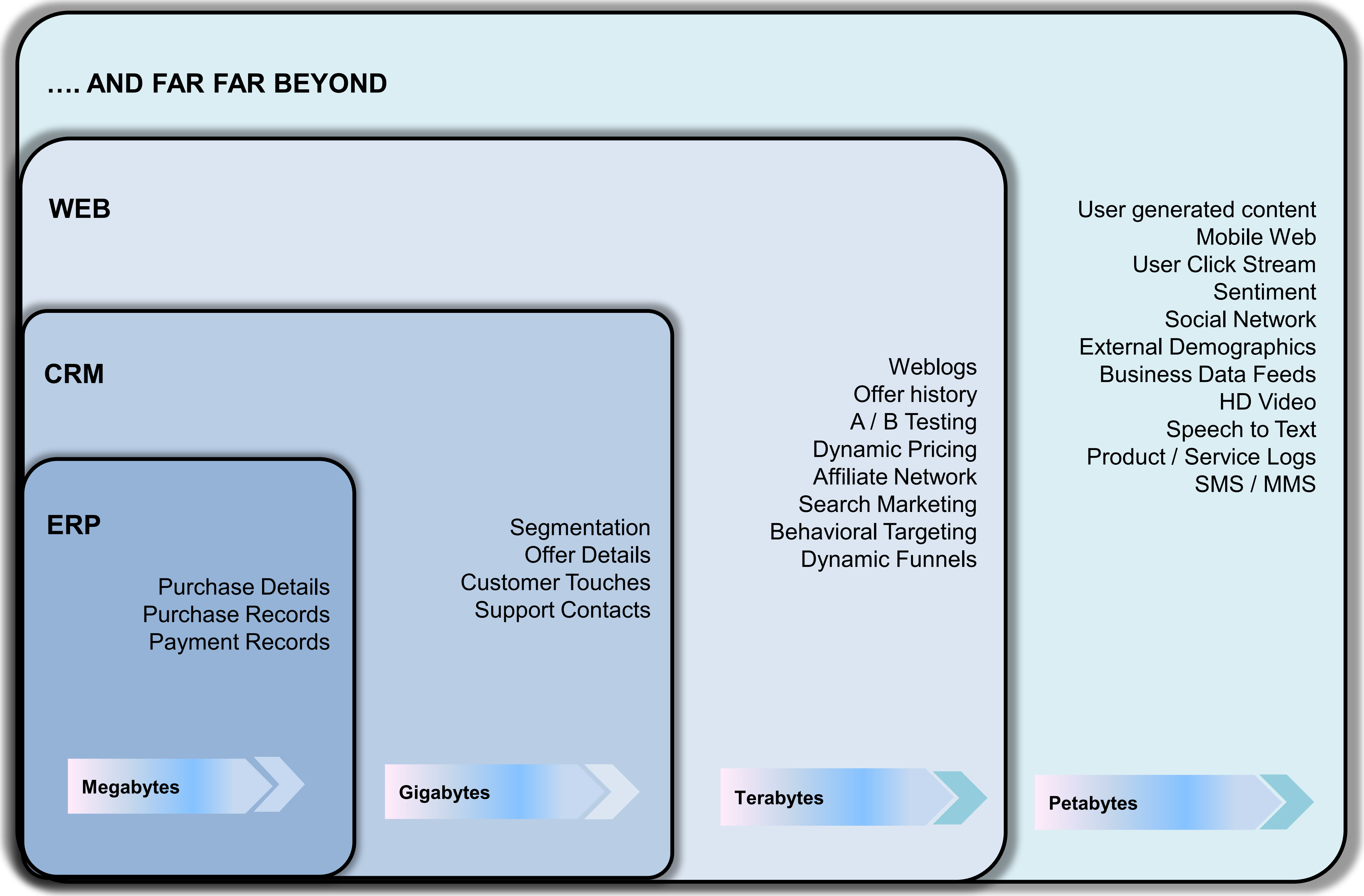
Agenda

- Introduction to Big Data
- Basics of Hadoop
- Hadoop MapReduce WordCount Demo
- Hadoop Ecosystem landscape
- Basics of Pig and Pig Latin
- Pig WordCount Demo
- Pig vs SQL and Pig vs Hive
- Visualization of Pig MapReduce Jobs with Twitter Ambrose

Pre-requisites

- Basic understanding of Hadoop, HDFS and MapReduce.
- Laptop with VMware Player or Oracle VirtualBox installed.
- Please copy the VMware image of 64 bit Ubuntu Server 12.04 distributed in the USB flash drive.
- Uncompress the VMware image and launch the image using VMware Player / Virtual Box.
- Login to the VM with the credentials:
  - hduser / hduser
- Check if the environment variables HADOOP_HOME, PIG_HOME, etc are set.

Introduction to Big Data



…. AND FAR FAR BEYOND

User generated content
Mobile Web
User Click Stream
Sentiment
Social Network
External Demographics
Business Data Feeds
HD Video
Speech to Text
Product / Service Logs
SMS / MMS

Petabytes

WEB

Weblogs
Offer history
A / B Testing
Dynamic Pricing
Affiliate Network
Search Marketing
Behavioral Targeting
Dynamic Funnels

Terabytes

CRM

Segmentation
Offer Details
Customer Touches
Support Contacts

Gigabytes

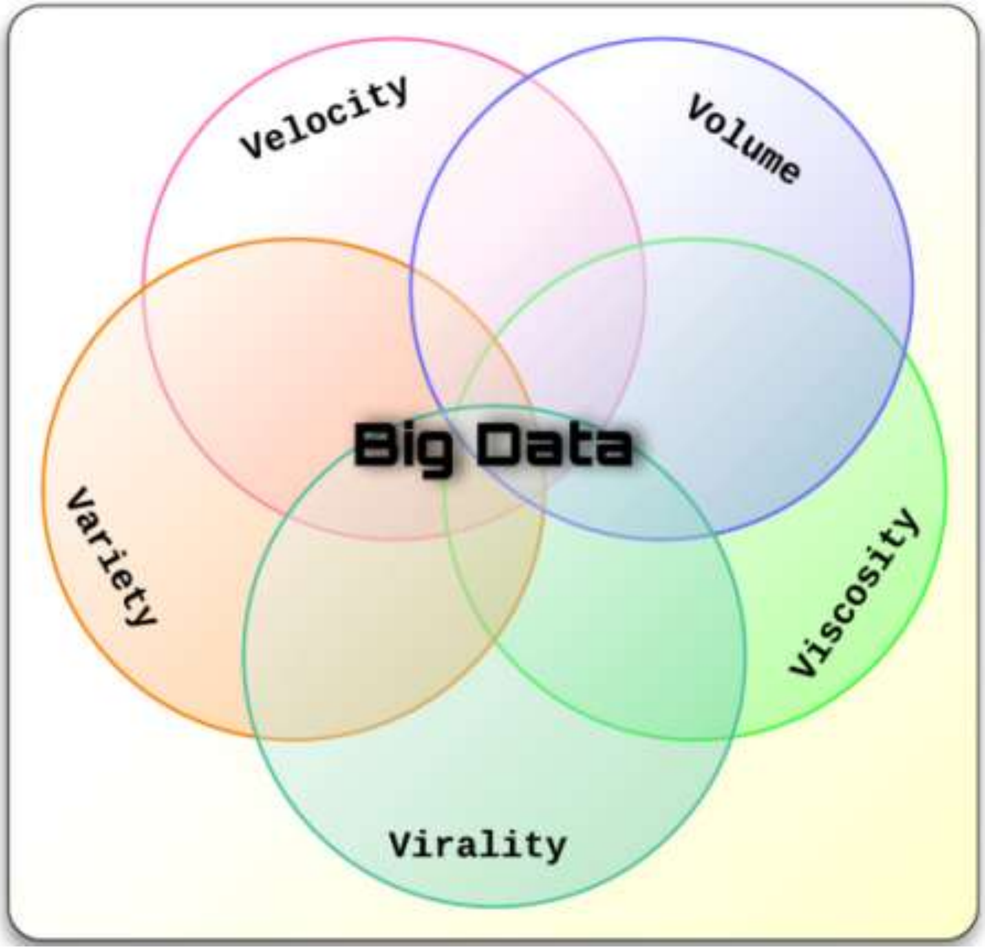
ERP

Purchase Details
Purchase Records
Payment Records

Megabytes

Source: http://datameer.com

Introduction to Big Data



Source: http://blog.softwareinsider.org/2012/02/27/mondays-musings-beyond-the-three-vs-of-big-data-viscosity-and-virality/

Big Data Analysis

❑ RDBMS (scalability)

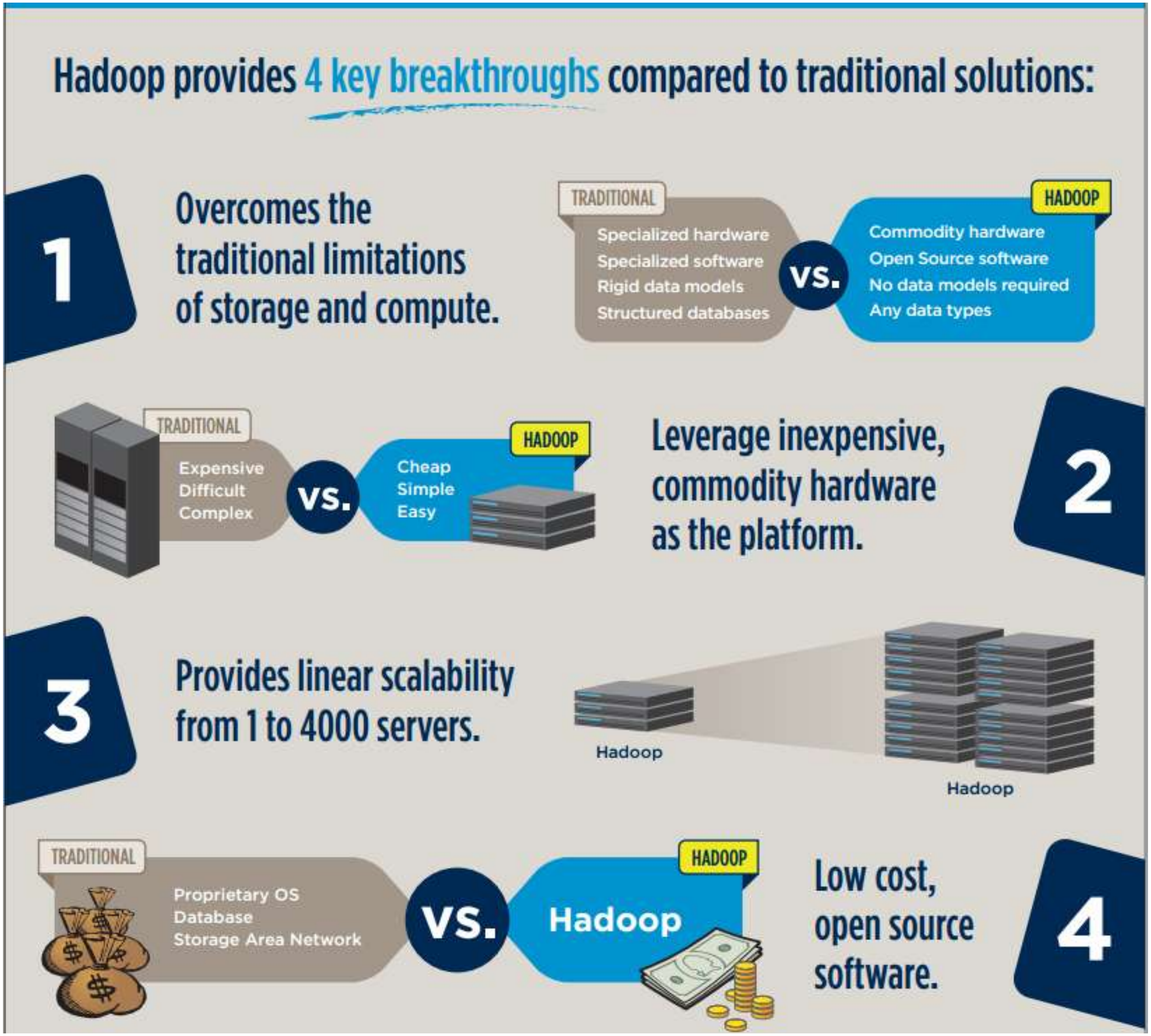
❑ Parallel RDBMS (expensive)

❑ Programming Language (too complex)

Hadoop comes to the rescue

hadoop

Why Hadoop?



Hadoop provides 4 key breakthroughs compared to traditional solutions:

1 Overcomes the traditional limitations of storage and compute.

TRADITIONAL		HADOOP
Specialized hardware Specialized software Rigid data models Structured databases	VS.	Commodity hardware Open Source software No data models required Any data types

2 Leverage inexpensive, commodity hardware as the platform.

TRADITIONAL		HADOOP
Expensive Difficult Complex	VS.	Cheap Simple Easy

3 Provides linear scalability from 1 to 4000 servers.

4 Low cost, open source software.

TRADITIONAL		HADOOP
Proprietary OS Database Storage Area Network	VS.	Hadoop

Source: http://datameer.com/pdf/WhyHadoop_HI.pdf

History of Hadoop

Scalable distributed file system for large distributed data-intensive applications

“The Google File System” by Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
http://research.google.com/archive/gfs.html

Programming model and an associated implementation for processing and generating large data sets`

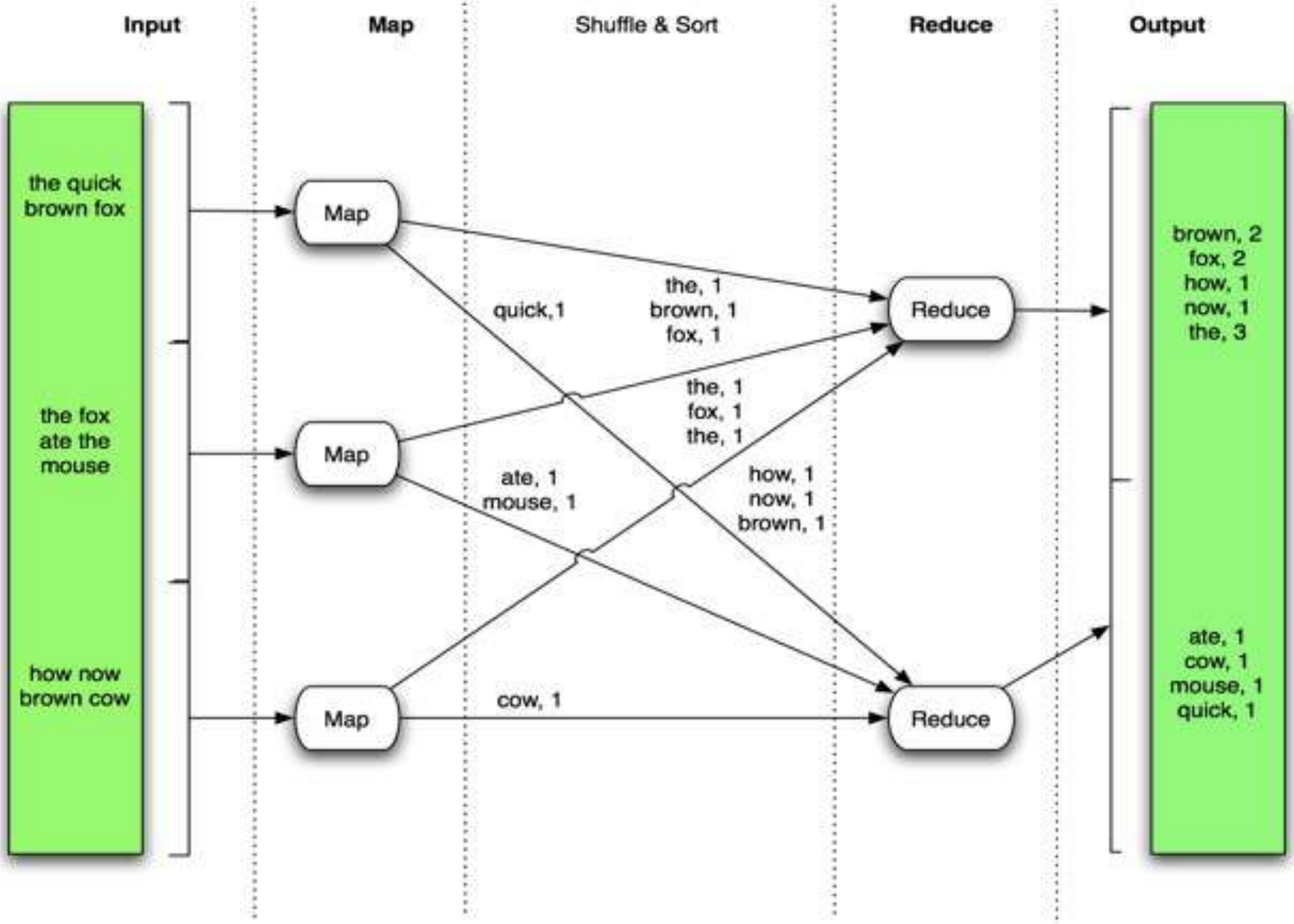
“MapReduce: Simplified Data Processing on Large Clusters” by Jeffrey Dean and Sanjay Ghemawat
http://research.google.com/archive/mapreduce.html

Introduction to Hadoop

- HDFS
  - Hadoop Distributed File System
  - A distributed, scalable, and portable filesystem written in Java for the Hadoop framework
  - Provides high-throughput access to application data.
  - Runs on large clusters of commodity machines
  - Is used to store large datasets.

- MapReduce
  - Distributed data processing model and execution environment that runs on large clusters of commodity machines
  - Also called MR.
  - Programs are inherently parallel.

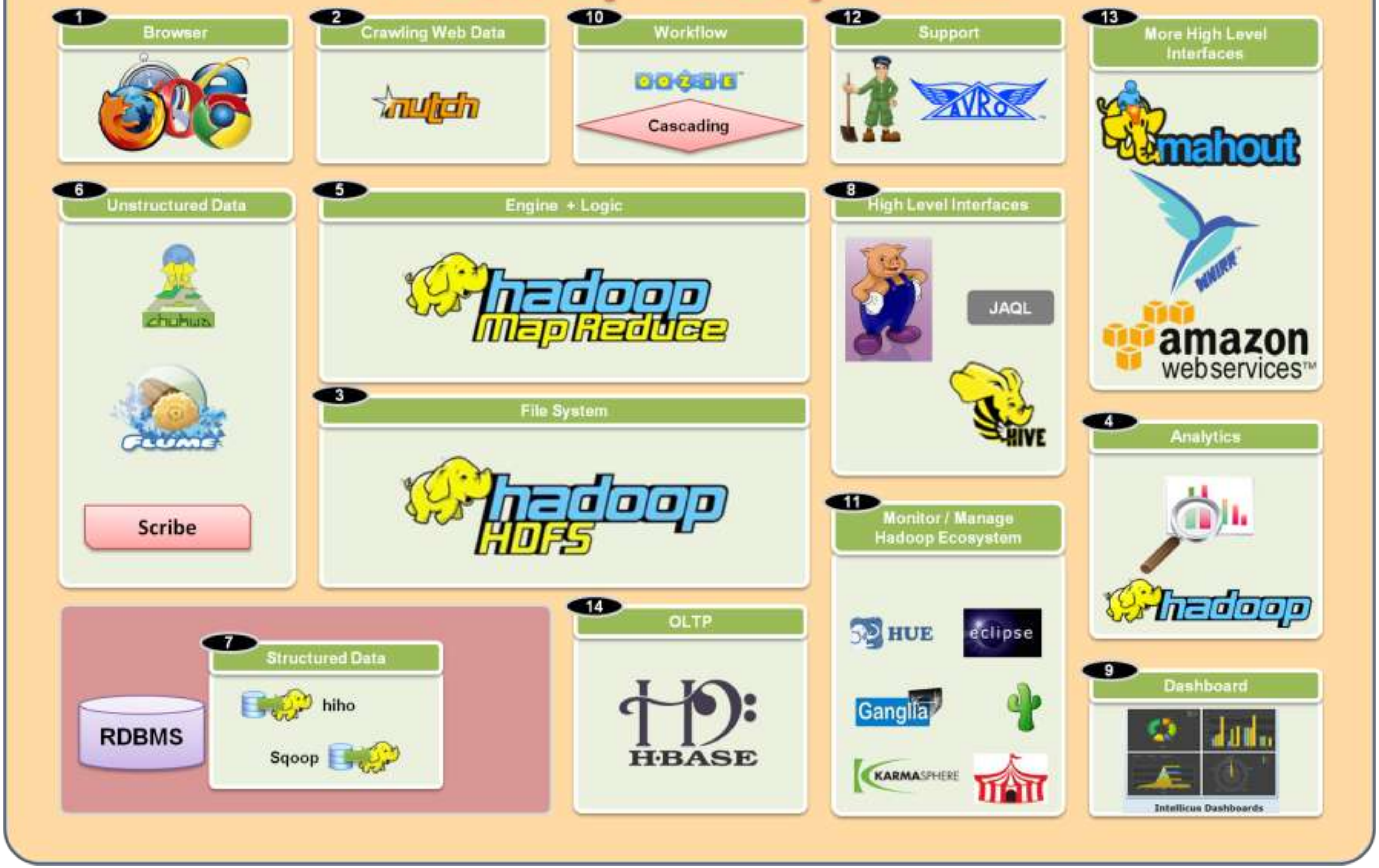
MapReduce



Source: http://www.slideshare.net/hadoop/practical-problem-solving-with-apache-hadoop-pig

Java MapReduce WordCount Example Demo

Hadoop Ecosystem



1 Browser

2 Crawling Web Data

10 Workflow

Cascading

12 Support

13 More High Level Interfaces

6 Unstructured Data

Scribe

5 Engine + Logic

8 High Level Interfaces

JAQL

3 File System

4 Analytics

11 Monitor / Manage Hadoop Ecosystem

14 OLTP

7 Structured Data

hiho

Sqoop

RDBMS

9 Dashboard

Source: http://indoos.wordpress.com/2010/08/16/hadoop-ecosystem-world-map/

Pig

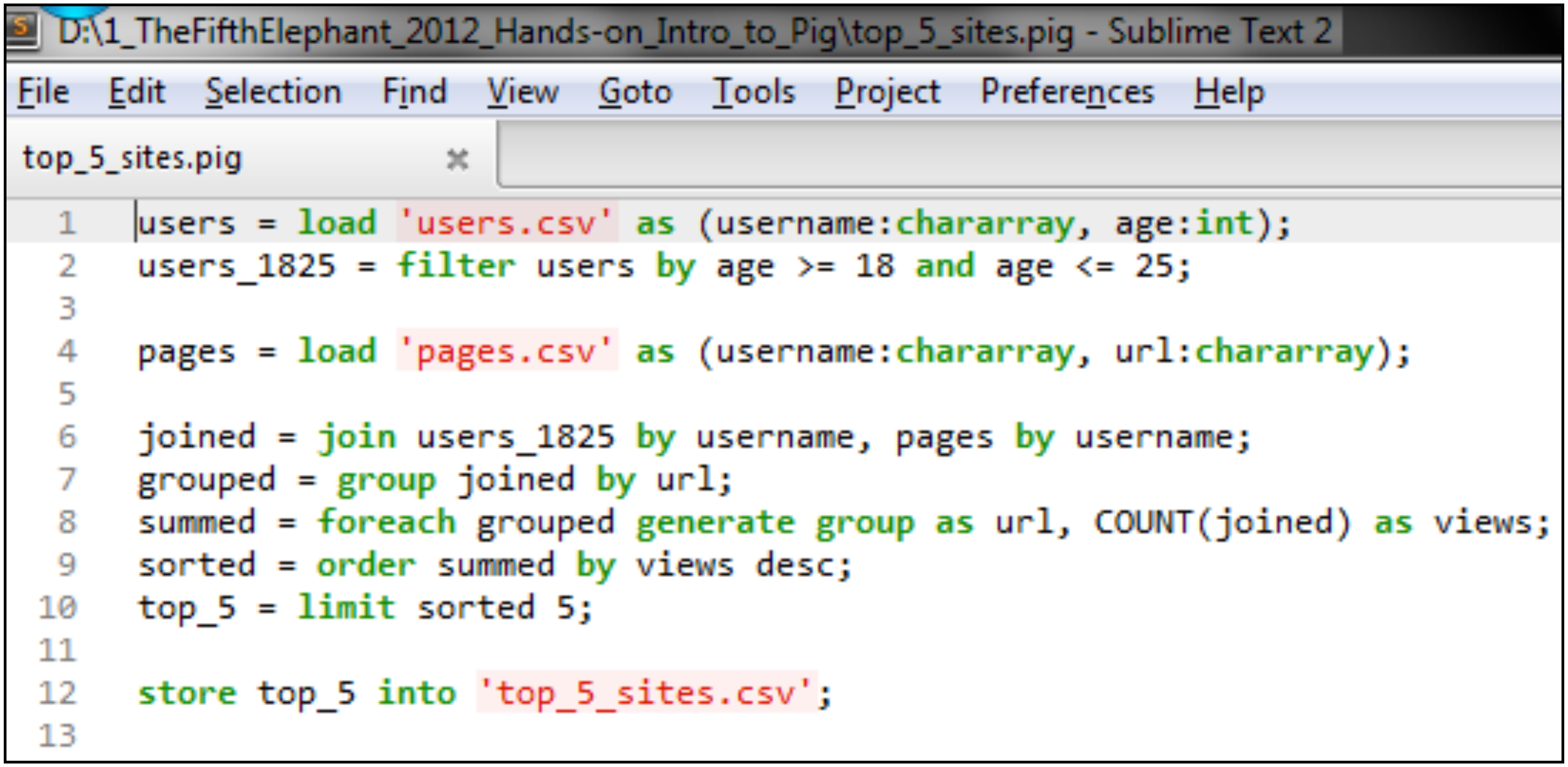
- **“Pig Latin: A Not-So-Foreign Language for Data Processing”**
  - Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, Andrew Tomkins (Yahoo! Research)
  - http://www.sigmod08.org/program_glance.shtml#sigmod_industrial_program
  - http://infolab.stanford.edu/~usriv/papers/pig-latin.pdf

Pig

- ❑ High level data flow language for exploring very large datasets.
- ❑ Provides an engine for executing data flows in parallel on Hadoop.
- ❑ Compiler that produces sequences of MapReduce programs
- ❑ Structure is amenable to substantial parallelization
- ❑ Operates on files in HDFS
- ❑ Metadata not required, but used when available

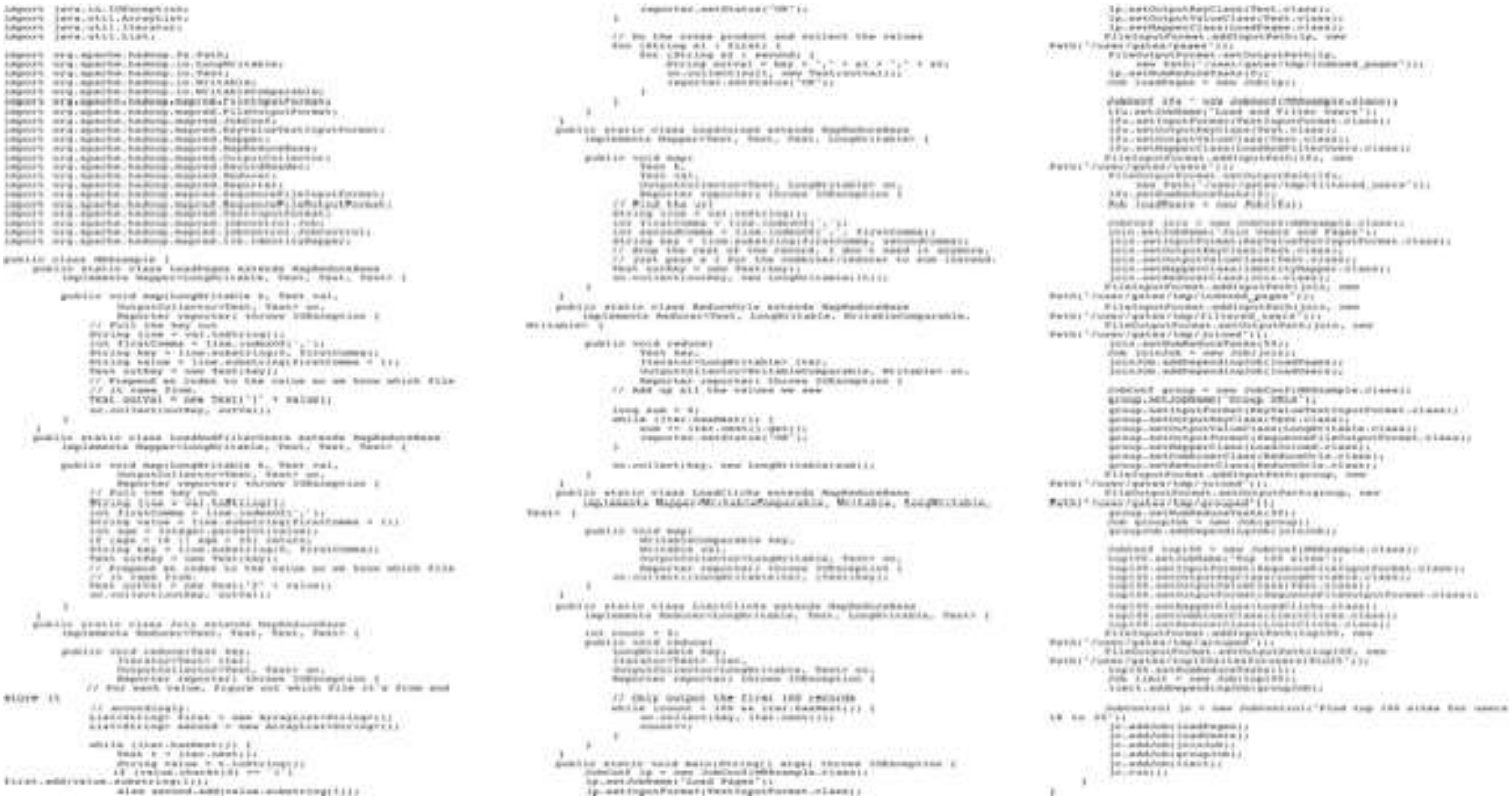
- ❑ Key Properties of Pig:
  - ➢ Ease of programming: Trivial to achieve parallel execution of simple and parallel data analysis tasks
  - ➢ Optimization opportunities: Allows the user to focus on semantics rather than efficiency
  - ➢ Extensibility: Users can create their own functions to do special-purpose processing

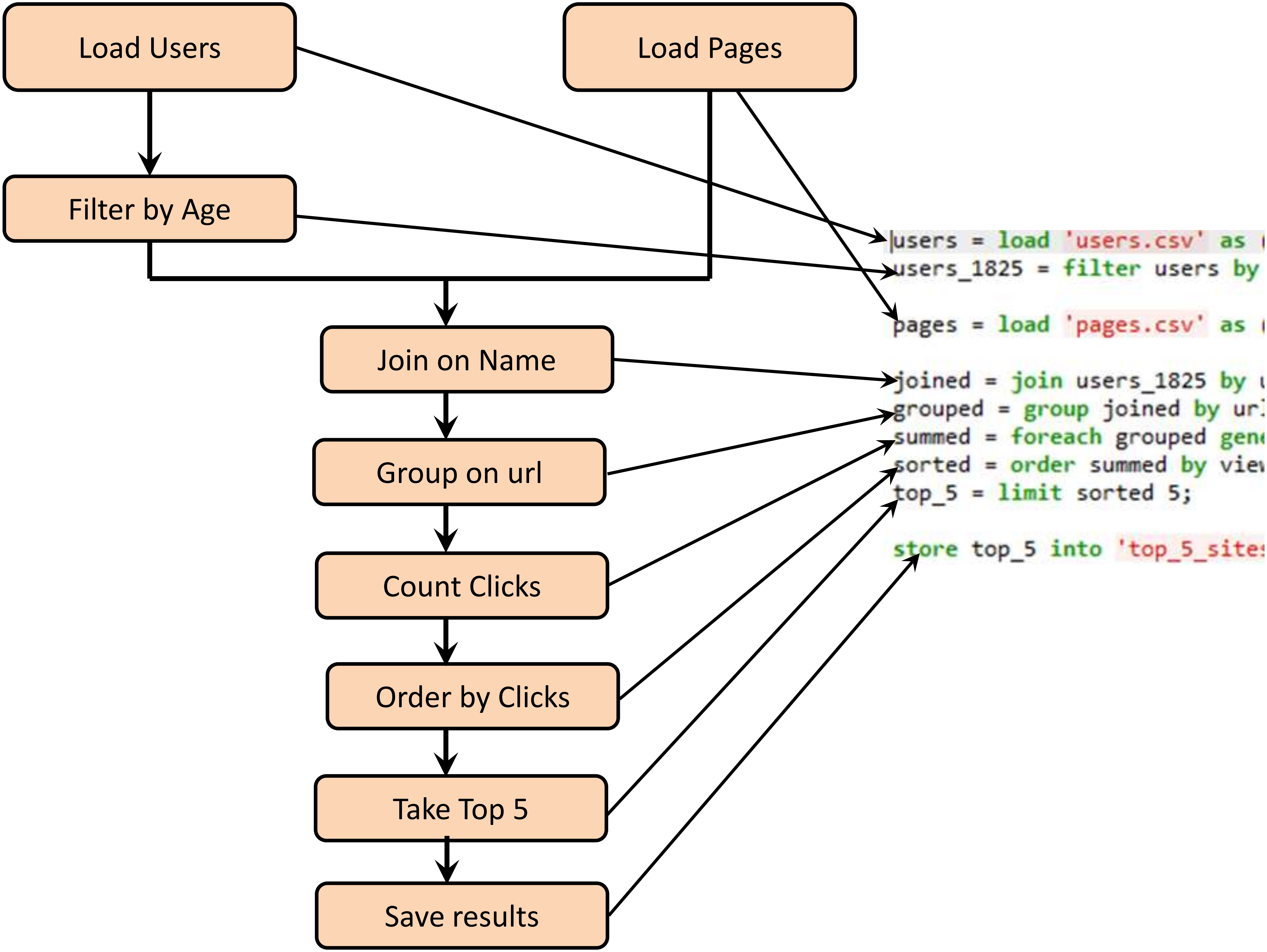
Why Pig?


D:\1_TheFifthElephant_2012_Hands-on_Intro_to_Pig\top_5_sites.pig - Sublime Text 2
File
Edit
Selection
Find
View
Goto
Tools
Project
Preferences
Help
top_5_sites.pig

```
users = load 'users.csv' as (username:chararray, age:int);
users_1825 = filter users by age >= 18 and age <= 25;

pages = load 'pages.csv' as (username:chararray, url:chararray);

joined = join users_1825 by username, pages by username;
grouped = group joined by url;
summed = foreach grouped generate group as url, COUNT(joined) as views;
sorted = order summed by views desc;
top_5 = limit sorted 5;

store top_5 into 'top_5_sites.csv';

```

Equivalent Java MapReduce Code





Load Users

Load Pages

Filter by Age

Join on Name

Group on url

Count Clicks

Order by Clicks

Take Top 5

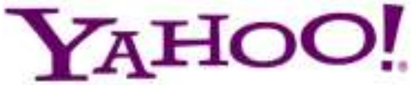
Save results

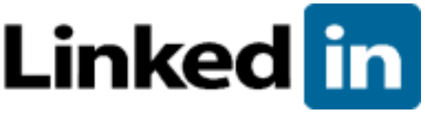
```
users = load 'users.csv' as
users_1825 = filter users by

pages = load 'pages.csv' as

joined = join users_1825 by
grouped = group joined by ur
summed = foreach grouped gen
sorted = order summed by vie
top_5 = limit sorted 5;

store top_5 into 'top_5_site
```

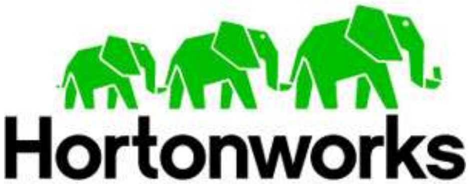
Pig vs Hadoop

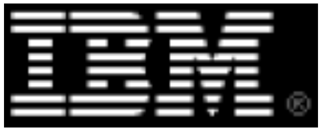
- 5% of the MR code.
- 5% of the MR development time.
- Within 25% of the MR execution time.
- Readable and reusable.
- Easy to learn DSL.
- Increases programmer productivity.
- No Java expertise required.
- Anyone [eg. BI folks] can trigger the Jobs.
- Insulates against Hadoop complexity
  - Version upgrades
  - Changes in Hadoop interfaces
  - JobConf configuration tuning
  - Job Chains

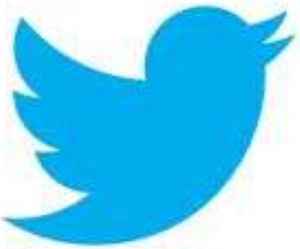
Committers of Pig

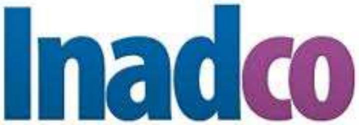


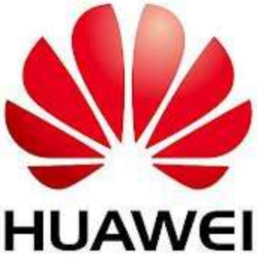


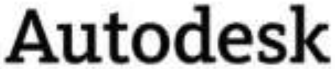


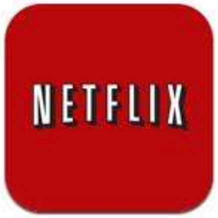












Source: http://pig.apache.org/whoweare.html

Who is using Pig?

Hortonworks
HUAWEI
IBM
YAHOO!
NOKIA
Connecting People
NETFLIX
Inadco
ChaCha
Aol.
amazon.com
cooliris
Linked in
Autodesk

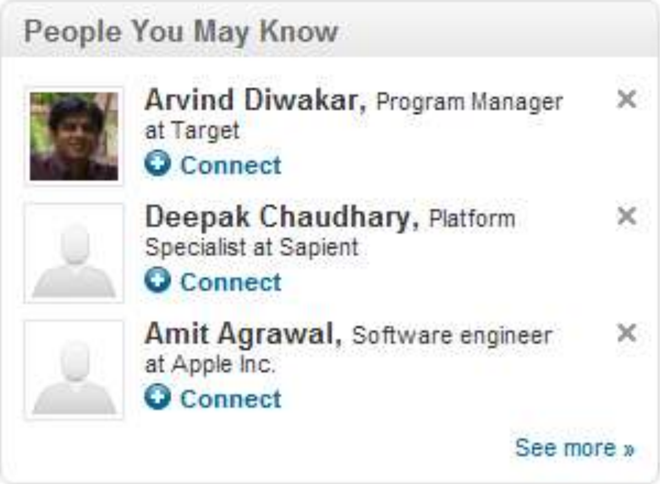
Source: http://wiki.apache.org/pig/PoweredBy

Pig use cases

- ❑ **Processing many Data Sources**
- ❑ **Data Analysis**
- ❑ **Text Processing**
  - ➢ **Structured**
  - ➢ **Semi-Structured**
- ❑ **ETL**
- ❑ **Machine Learning**
- ❑ **Advantage of Sampling in any use case**

Pig in real-world

LinkedIn



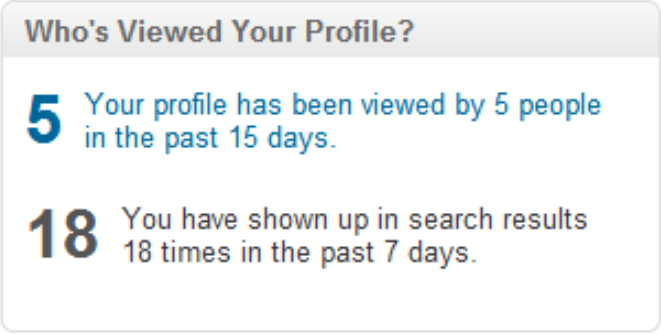
People You May Know

Arvind Diwakar, Program Manager at Target
Connect

Deepak Chaudhary, Platform Specialist at Sapient
Connect

Amit Agrawal, Software engineer at Apple Inc.
Connect

See more »

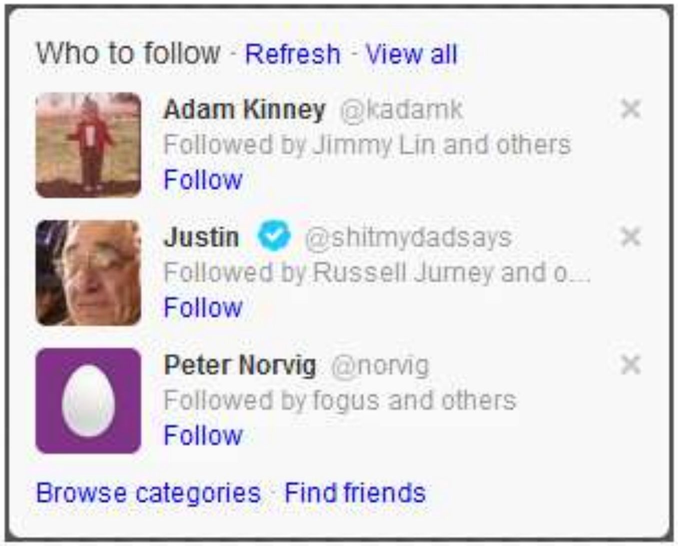


Who's Viewed Your Profile?

5 Your profile has been viewed by 5 people in the past 15 days.

18 You have shown up in search results 18 times in the past 7 days.

Twitter



Who to follow · Refresh · View all

Adam Kinney @kadamk
Followed by Jimmy Lin and others
Follow

Justin @shitmydadsays
Followed by Russell Jurney and o...
Follow

Peter Norvig @norvig
Followed by fogus and others
Follow

Browse categories · Find friends

Reporting, ETL, targeted emails & recommendations, spam analysis, ML

Components of Pig

- ❑ **Pig Latin**
  - ➢ **Submit a script directly**
- ❑ **Grunt**
  - ➢ **Pig Shell**
- ❑ **PigServer**
  - ➢ **Java Class similar to JDBC interface**

Pig Execution Modes

❑ Local Mode

- ➢ **Need access to a single machine**
- ➢ **All files are installed and run using your local host and file system**
- ➢ **Is invoked by using the** *`-x local`* **flag**
  - ▪ `pig -x local`

❑ MapReduce Mode

- ➢ **Mapreduce mode is the default mode**
- ➢ **Need access to a Hadoop cluster and HDFS installation.**
- ➢ **Can also be invoked by using the** *`-x mapreduce`* **flag or just** `pig`
  - ▪ `pig`
  - ▪ `pig -x mapreduce`

Pig Latin Statements

- ❑ Pig Latin Statements work with relations
  - ➢ Field is a piece of data.
    - ➢ John
  - ➢ Tuple is an ordered set of fields.
    - ➢ (John,18,4.0F)
  - ➢ Bag is a collection of tuples.
    - ➢ (1,{(1,2,3)})
  - ➢ Relation is a bag

Pig Simple Datatypes

Simple Type	Description	Example
int	Signed 32-bit integer	10
long	Signed 64-bit integer	Data: 10L or 10l Display: 10L
float	32-bit floating point	Data: 10.5F or 10.5f or 10.5e2f or 10.5E2F Display: 10.5F or 1050.0F
double	64-bit floating point	Data: 10.5 or 10.5e2 or 10.5E2 Display: 10.5 or 1050.0
chararray	Character array (string) in Unicode UTF-8 format	hello world
bytearray	Byte array (blob)	
boolean	boolean	true/false (case insensitive)

Pig Complex Datatypes

Type	Description	Example
tuple	An ordered set of fields.	(19,2)
bag	An collection of tuples.	{(19,2), (18,1)}
map	A set of key value pairs.	[open#apache]

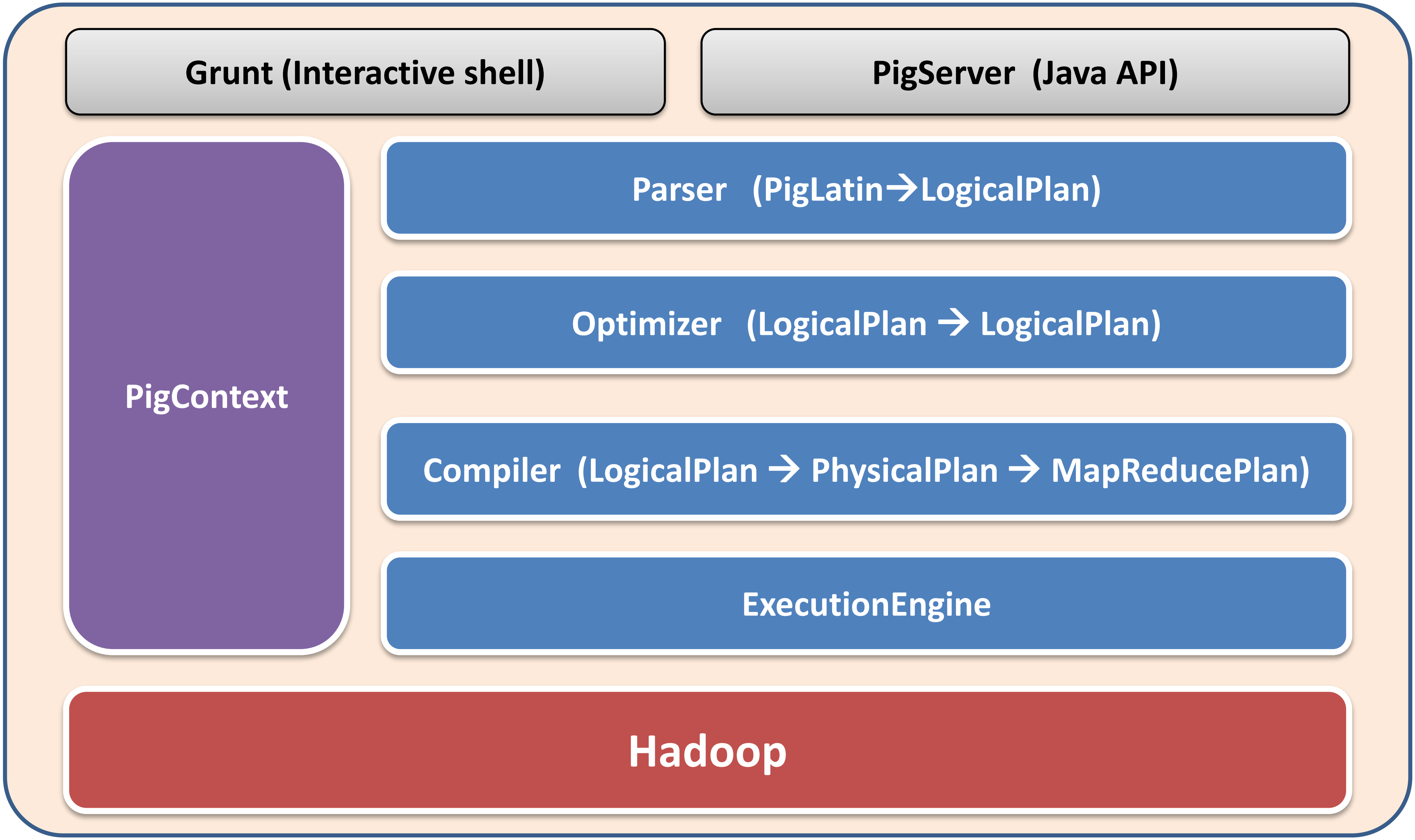
Pig Commands

Statement	Description
Load	Read data from the file system
Store	Write data to the file system
Dump	Write output to stdout
Foreach	Apply expression to each record and generate one or more records
Filter	Apply predicate to each record and remove records where false
Group / Cogroup	Collect records with the same key from one or more inputs
Join	Join two or more inputs based on a key
Order	Sort records based on a Key
Distinct	Remove duplicate records
Union	Merge two datasets
Limit	Limit the number of records
Split	Split data into 2 or more sets, based on filter conditions

Pig Diagnostic Operators

Statement	Description
Describe	Returns the schema of the relation
Dump	Dumps the results to the screen
Explain	Displays execution plans.
Illustrate	Displays a step-by-step execution of a sequence of statements

Architecture of Pig



Grunt (Interactive shell)

PigServer (Java API)

PigContext

Parser (PigLatin→LogicalPlan)

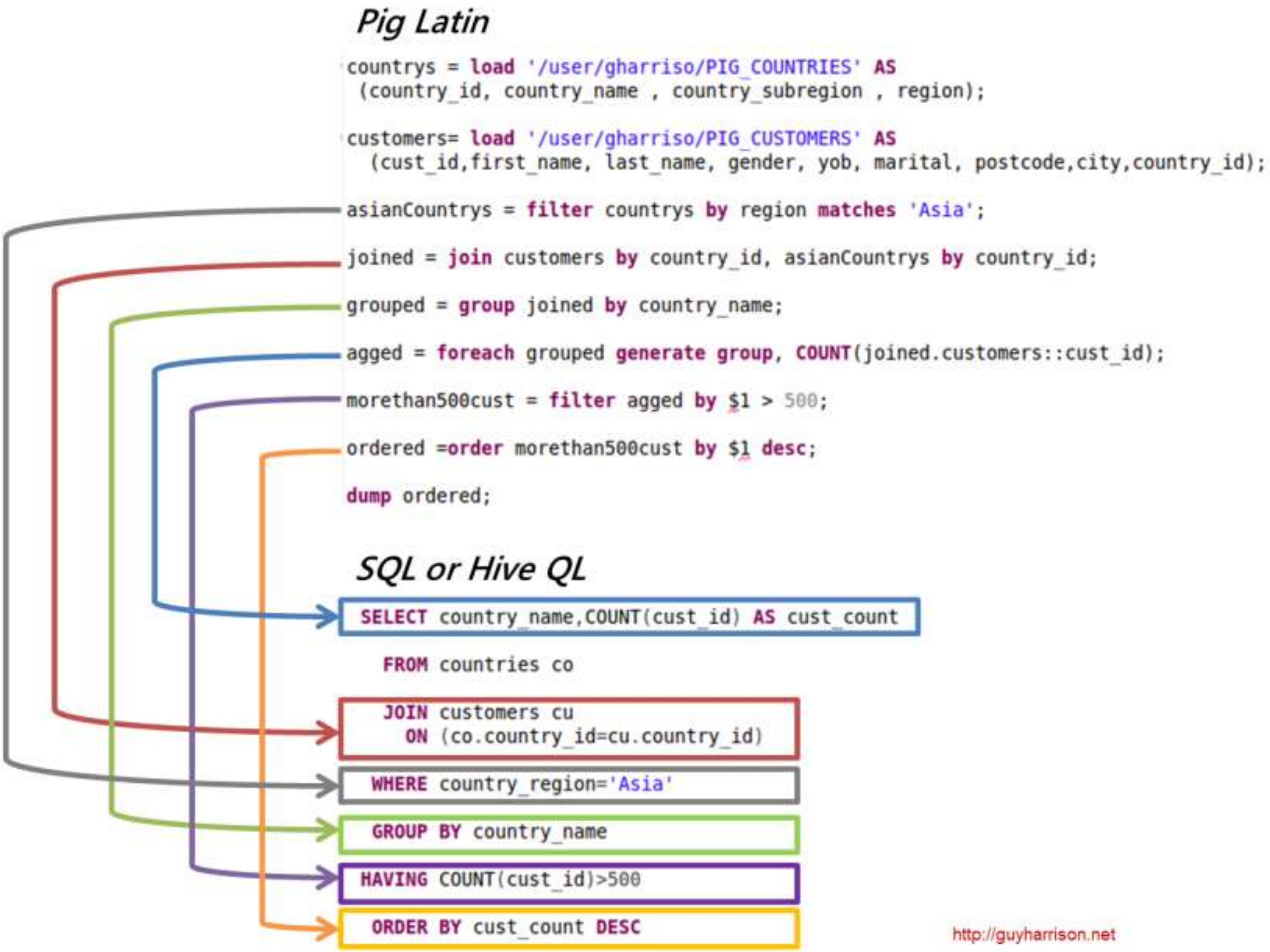
Optimizer (LogicalPlan → LogicalPlan)

Compiler (LogicalPlan → PhysicalPlan → MapReducePlan)

ExecutionEngine

Hadoop

Pig Latin vs SQL



Pig Latin

```
countrys = load '/user/gharriso/PIG_COUNTRIES' AS
 (country_id, country_name , country_subregion , region);

customers= load '/user/gharriso/PIG_CUSTOMERS' AS
  (cust_id,first_name, last_name, gender, yob, marital, postcode,city,country_id);

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
```

http://guyharrison.net

Pig vs SQL

Pig	SQL
Dataflow	Declarative
Nested relational data model	Flat relational data model
Optional Schema	Schema is required
Scan-centric workloads	OLTP + OLAP workloads
Limited query optimization	Significant opportunity for query optimization

Source: http://www.slideshare.net/hadoop/practical-problem-solving-with-apache-hadoop-pig

Hive Demo

Pig vs Hive

Feature	Pig	Hive
Language	PigLatin	SQL-like
Schemas / Types	Yes (implicit)	Yes (explicit)
Partitions	No	Yes
Server	No	Optional (Thrift)
User Defined Functions (UDF)	Yes (Java, Python, Ruby, etc)	Yes (Java)
Custom Serializer/Deserializer	Yes	Yes
DFS Direct Access	Yes (explicit)	Yes (implicit)
Join/Order/Sort	Yes	Yes
Shell	Yes	Yes
Streaming	Yes	Yes
Web Interface	No	Yes
JDBC/ODBC	No	Yes (limited)

Source:http://www.larsgeorge.com/2009/10/hive-vs-pig.html

Storage Options in Pig

- ❑ **HDFS**
  - ➢ Plain Text
  - ➢ Binary format
  - ➢ Customized format (XML, JSON, Protobuf, Thrift, etc)

- ❑ **RDBMS** (DBStorage)

- ❑ **Cassandra** (CassandraStorage)

- ❑ **HBase** (HBaseStorage)

- ❑ **Avro** (AvroStorage)

Visualization of Pig MapReduce Jobs

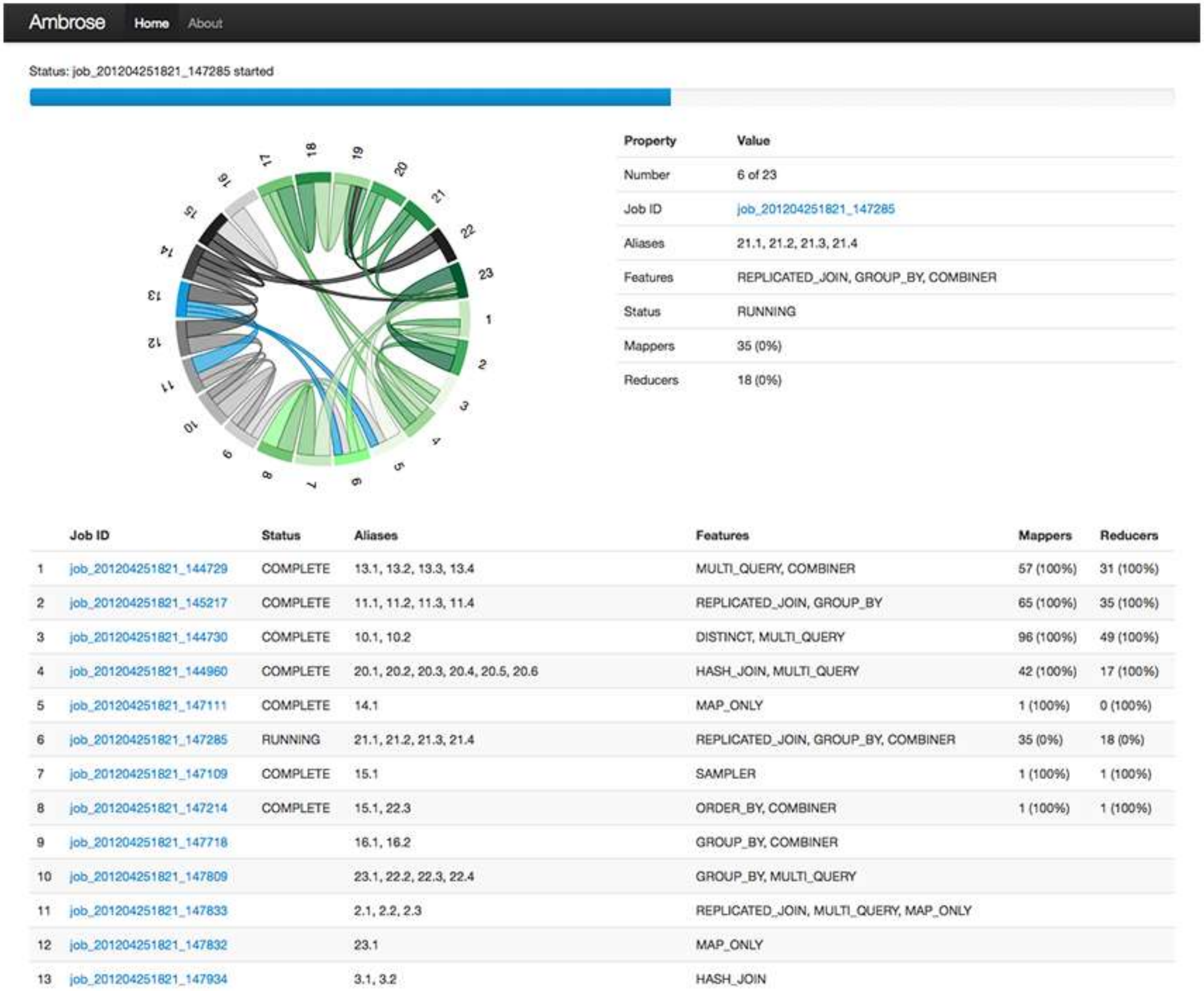
- **Twitter Ambrose:** https://github.com/twitter/ambrose
  - Platform for visualization and real-time monitoring of MapReduce data workflows
  - Presents a global view of all the MapReduce jobs derived from the workflow after planning and optimization

- **Ambrose provides the following in a web UI:**
  - A chord diagram to visualize job dependencies and current state
  - A table view of all the associated jobs, along with their current state
  - A highlight view of the currently running jobs
  - An overall script progress bar

- **Ambrose is built using:**
  - D3.js
  - Bootstrap

- **Supported Runtimes:** Designed to support any Hadoop workflow runtime
  - **Currently supports Pig MR Jobs**
  - Future work would include Cascading, Scalding, Cascalog and Hive

Twitter Ambrose



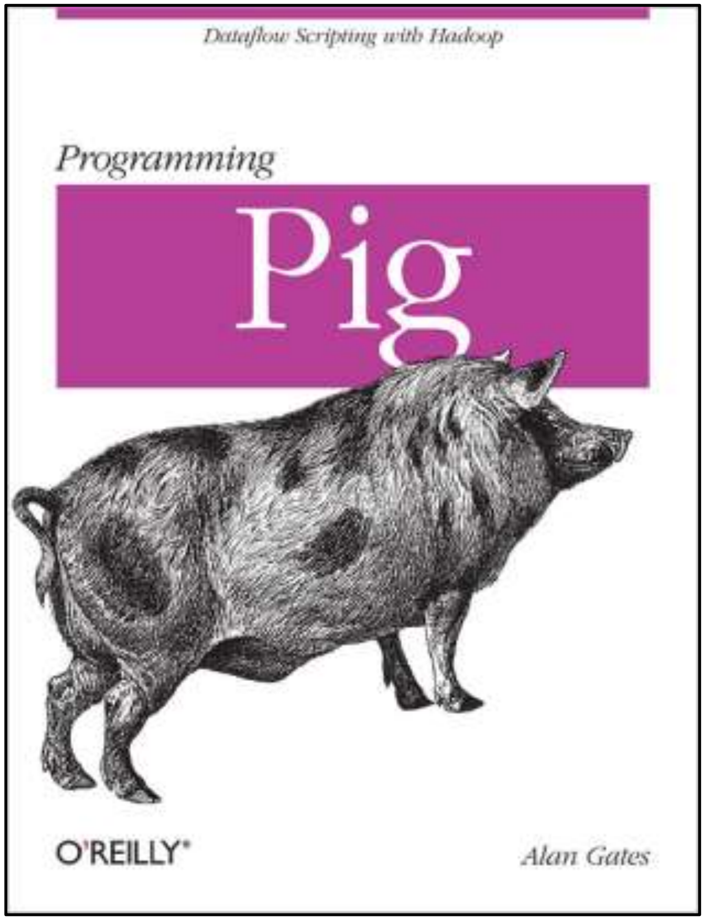
Status: job_201204251821_147285 started

Property	Value
Number	6 of 23
Job ID	job_201204251821_147285
Aliases	21.1, 21.2, 21.3, 21.4
Features	REPLICATED_JOIN, GROUP_BY, COMBINER
Status	RUNNING
Mappers	35 (0%)
Reducers	18 (0%)

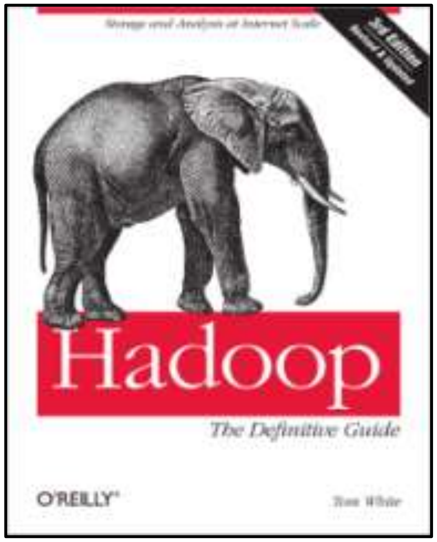
	Job ID	Status	Aliases	Features	Mappers	Reducers
1	job_201204251821_144729	COMPLETE	13.1, 13.2, 13.3, 13.4	MULTI_QUERY, COMBINER	57 (100%)	31 (100%)
2	job_201204251821_145217	COMPLETE	11.1, 11.2, 11.3, 11.4	REPLICATED_JOIN, GROUP_BY	65 (100%)	35 (100%)
3	job_201204251821_144730	COMPLETE	10.1, 10.2	DISTINCT, MULTI_QUERY	96 (100%)	49 (100%)
4	job_201204251821_144960	COMPLETE	20.1, 20.2, 20.3, 20.4, 20.5, 20.6	HASH_JOIN, MULTI_QUERY	42 (100%)	17 (100%)
5	job_201204251821_147111	COMPLETE	14.1	MAP_ONLY	1 (100%)	0 (100%)
6	job_201204251821_147285	RUNNING	21.1, 21.2, 21.3, 21.4	REPLICATED_JOIN, GROUP_BY, COMBINER	35 (0%)	18 (0%)
7	job_201204251821_147109	COMPLETE	15.1	SAMPLER	1 (100%)	1 (100%)
8	job_201204251821_147214	COMPLETE	15.1, 22.3	ORDER_BY, COMBINER	1 (100%)	1 (100%)
9	job_201204251821_147718		16.1, 16.2	GROUP_BY, COMBINER		
10	job_201204251821_147809		23.1, 22.2, 22.3, 22.4	GROUP_BY, MULTI_QUERY		
11	job_201204251821_147833		2.1, 2.2, 2.3	REPLICATED_JOIN, MULTI_QUERY, MAP_ONLY		
12	job_201204251821_147832		23.1	MAP_ONLY		
13	job_201204251821_147934		3.1, 3.2	HASH_JOIN		

Twitter Ambrose
Demo

Books

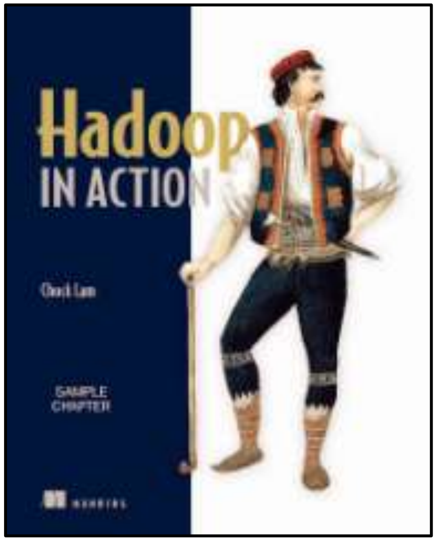


http://amzn.com/1449302645



http://amzn.com/1449311520

Chapter:11 “*Pig”*



http://amzn.com/1935182196

Chapter:10 “*Programming with Pig”*

Further Study & Blog-roll

- ❑ Online documentation: http://pig.apache.org
- ❑ Pig Confluence: https://cwiki.apache.org/confluence/display/PIG/Index
- ❑ Online Tutorials:
  - ➢ Cloudera Training, http://www.cloudera.com/resource/introduction-to-apache-pig/
  - ➢ Yahoo Training, http://developer.yahoo.com/hadoop/tutorial/pigtutorial.html
  - ➢ Using Pig on EC2:

    http://developer.amazonwebservices.com/connect/entry.jspa?externalID=2728
- ❑ Join the mailing lists:
  - ➢ Pig User Mailing list, user@pig.apache.org
  - ➢ Pig Developer Mailing list, dev@pig.apache.org

Trainings and Certifications

- Cloudera: http://university.cloudera.com/training/apache_hive_and_pig/hive_and_pig.html
- Hortonworks: http://hortonworks.com/hadoop-training/hadoop-training-for-developers/

Questions

Thank You