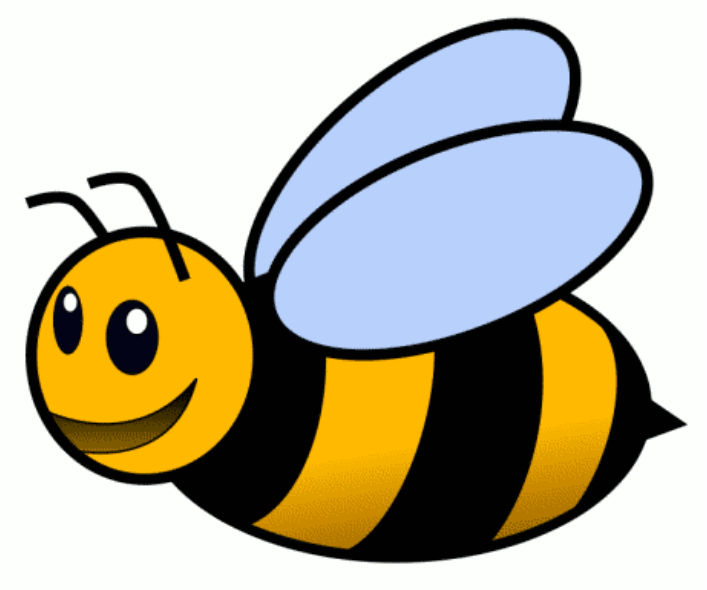
Jive with Hive



Allan Mitchell

- Joint author on 2005/2008 SSIS Book by Wrox
- Websites
  - www.CopperBlueConsulting.com
- Specialise in Data and Process Integration
- Microsoft SQL Server MVP
- Twitter: allanSQLIS
- E: allan.mitchell@Copper-Blue.com

Agenda

Hive solves the business problem of analyzing large amounts of data

- A brief summary of Hadoop and Big Data
- What is the purpose of Hive?
- Why Hive?
- A history of Hive
- What are Hive's constituents

Agenda

Hive solves the business problem of analyzing large amounts of data

- **A brief summary of Hadoop and Big Data**
- What is the purpose of Hive?
- Why Hive?
- A history of Hive
- What are Hive's constituents

What is Big Data

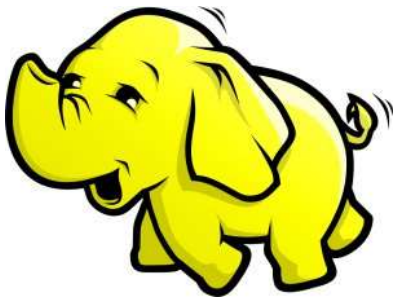
- Traditionally:
  - Physics Experiments, Sensor data, Satellite data, …
- Now:
  - Operational Logs
  - Customer behavior
  - Social interactions online
  - …
- From Terabytes in the 1990s over Petabytes today to Zetabytes in the future

What is Big Data

"When you have to innovate to collect, store, organize, analyse and share it"

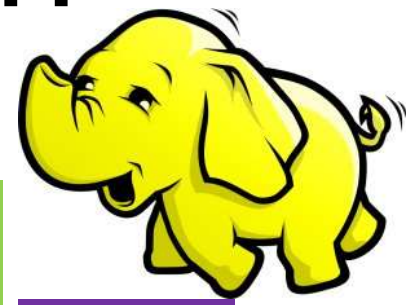
-Werner Vogels Amazon CTO

What is Hadoop?



"Flexible and Available Architecture for Large Scale computation and data processing on a network of highly available commodity hardware."

HDInsight Ecosystem



Metadata (HCatalog)

Graph (Pegasus)

Stats processing (RHadoop)

P D W

NoSQL Database (HBase)

Scripting (Pig)

Query (Hive)

Machine Learning (Mahout)

Data Integration (ODBC / SQOOP/ REST)

Business Intelligence (Excel, PowerView..)

Distributed Processing (Map Reduce)

Distributed Storage (HDFS)

(Azure Data Marketplace)

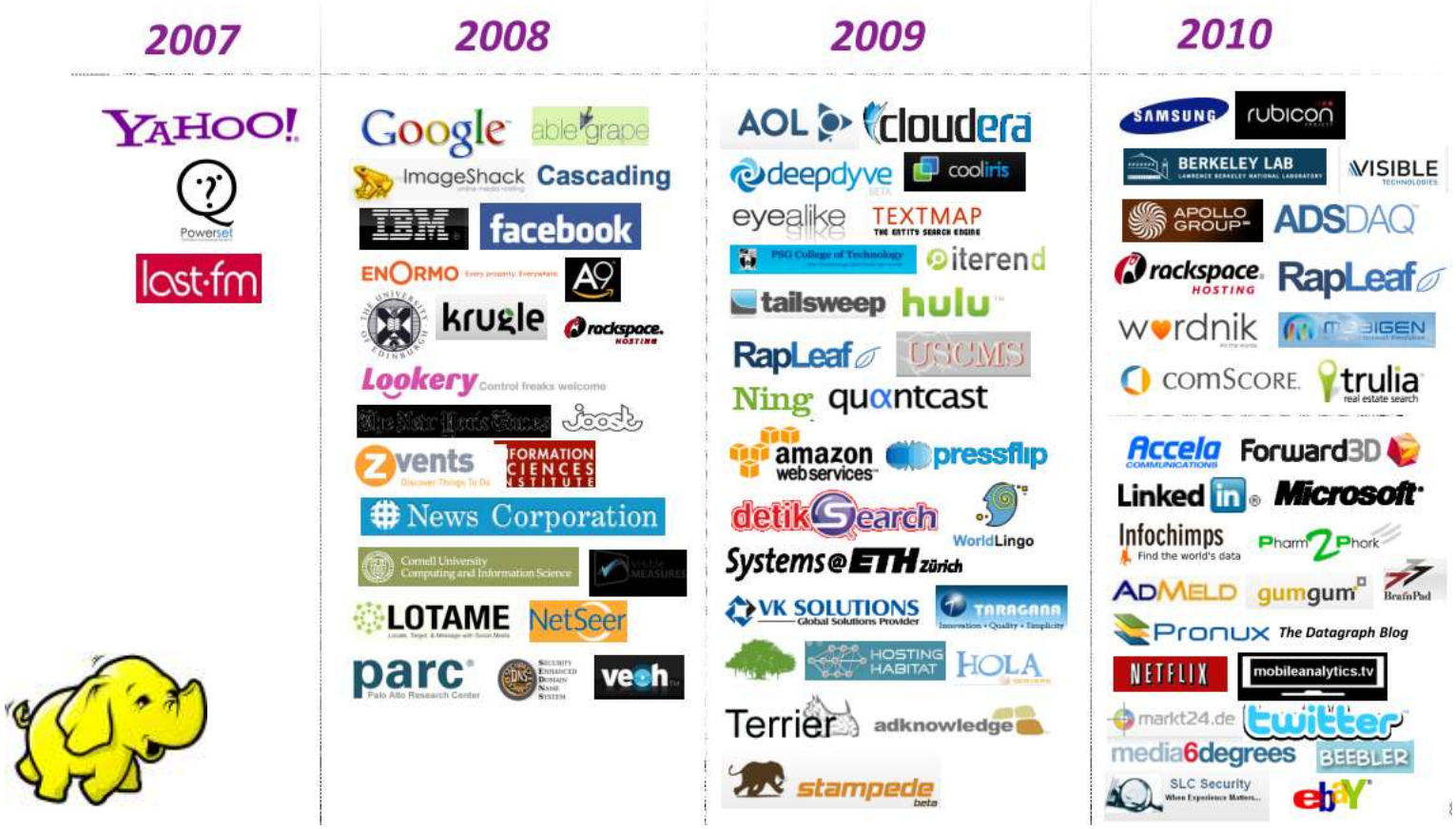
Windows Azure Storage

AD, System Center

HDInsight

- Hadoop
- Collaboration with Hortonworks
- Sandbox Download – Single node cluster
- Azure offering
- HDP 1.3 for Windows – multi-node cluster

Hadoop's Lineage



* Resource: Kerberos Konference (Yahoo) – 2010

Hadoop Key Terms

- HDFS – Distributed, Fault Tolerance File System
- MapReduce – Parallel Data Processing Framework
- Hive – Query Framework (Like SQL)
- Pig - Query Scripting Tool
- HBASE – Real Time access to Big Data

Agenda

Hive solves the business problem of analyzing large amounts of data

- A brief summary of Hadoop
- **What is the purpose of Hive?**
- Why Hive?
- A history of Hive
- What are Hive's constituents

What is the purpose of Hive?

Hive is a solution to a business problem:

How do you analyze large amounts of data?

Data Scientists want to study data

Communicate with the data

Businesses want to reap benefits of data

Results that make sense of the data


Self Service BI
Microsoft SQL Server 2012
Analysers
Data Analysis
Data Preparation
Data Visualisation
HIVE
Developers
hadoop
Data Management
Data Integration
Data Collection
Admins
Data Sources
Unstructured Data Sources
Docs
Emails
Videos
Semi-Structured Data Sources
JSON
Logs
Social
Structured Data Sources
LOB
CRM
ERP

What is the purpose of Hive?

Hive is a data warehousing system for Hadoop

To meet the needs of businesses, data scientists, analysts and BI professionals

Data, Summarized

Fit a structure onto data

Data, Analyzed

Analysis of Large Datasets stored in Hadoop File Systems
SQL-Like language called HiveQL
Custom mappers and reduces when HiveQL isn't enough

Agenda

- Hive solves the business problem of analyzing large amounts of data

- What is the purpose of Hive?
- **Why Hive?**
- A history of Hive
- What are Hive's constituents

Why Hive?

Can't Hadoop be used to solve these problems?

Why is there a need for Hive?

Writing MR jobs in Java can be difficult

You don't know it's wrong until it's fallen over!

Joining Large Datasets can be difficult

Learning Curve

Ordering Datasets requires being a Ninja

Agenda

- Hive solves the business problem of analyzing large amounts of data

- What is the purpose of Hive?
- Why Hive?
- **A history of Hive**
- What are Hive's constituents

Hive History

Hive History

What can Hive offer you?

Hive can help with a range of business problems:

- Log Processing
- Predictive Modelling
- Hypothesis testing
- And Business Intelligence

Hive is not a replacement for SQL

So don’t throw out your SQL Server instances!

- Hive is for processing large data sets that may span hundreds, or even thousands, of machines
- Hive has a high overhead for starting a job. It translates queries to MR so it takes time
- Hive does not cache data, like SQL Server
- Hive performance tuning is mainly Hadoop performance tuning
- Similarity of the query engine, but different architectures for different purposes

Agenda

Hive solves the business problem of analyzing large amounts of data

- What is the purpose of Hive?
- Why Hive?
- A history of Hive
- What are Hive's constituents?
  - **Hive as a SQL-like Language Query Tool**
  - **Hive as a Translation Tool**
  - **Hive as a Structuring Tool**

HiveQL

Hive QL is a SQL-like language

It outputs naturally occurring groups for further analysis

Easy Data Summarization

Large Datasets, summarized

Fit a structure onto data

Analysis of Large Datasets stored in Hadoop file systems

SQL-Like language called HiveQL

Custom mappers and reduces when HiveQL isn’t enough

HiveQL Queries like SQL Queries?

Similarities in Syntax and Features

Similar features

SELECT

FROM

WHERE

GROUP BY / HAVING

Table Aliases

Computed Columns

HiveQL Queries like SQL Queries?

Similarities in Syntax and Features

Similar features

Aggregate Functions
Nested Select
CASE
LIKE / RLIKE
JOIN
ORDER BY / SORT BY

How does Hive work?

Hive as a structuring Tool

Creates a schema around the data
Tables stored in Directories

Hive Tables

Rows and columns, like SQL tables

Hive Metastore

Namespace with a set of tables
Holds table definitions
- Physical Layout
- Column Types
- Partition Information

Hive DEMO

Hive – Create a Table v2

```
CREATE EXTERNAL TABLE Ext
        (
        Exch string,
        Symbol string,
        date string,
        val float
        )
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ‘,‘
LOCATION ‘asv:///inputfiles/’;
```

Hive – Create a Table v3

```
INSERT OVERWRITE TABLE
        SomeTable
SELECT
        pv_users.gender,
        count(DISTINCT pv_users.userid),
        count(*),
        sum(DISTINCT pv_users.userid)
FROM
        ADifferentTable
GROUP BY
        pv_users.gender;
```

Hive – Create a Table v4

```
CREATE TABLE
        SomeTable
LIKE ADifferentTable
```

Hive – Create a Table v5

```
CREATE TABLE SomeTable
AS
SELECT * FROM AnotherTable;
```

Notes about creating tables

- INTERNAL
  - Means Hadoop manages the whole deal
  - Drop the table then the data goes too
  - Useful for temporary objects
  - Cannot specify INTERNAL on CREATE statement
  - Default
- EXTERNAL
  - Hadoop manages the metadata
  - Drop just drops the metadata not the data
  - Must use EXTERNAL keyword

Hive – Joining Tables

```
SELECT a.val, b.val, c.val FROM a
JOIN b ON (a.key = b.key1) JOIN c ON (c.key = b.key2)
```

```
SELECT a.val, b.val FROM a LEFT OUTER JOIN b ON (a.key=b.key)
WHERE a.ds='2009-07-07' AND b.ds='2009-07-07‘
```

```
SELECT a.key, a.val FROM a LEFT SEMI JOIN b on (a.key = b.key)
```

Hive – Sampling

```
SELECT * FROM source TABLESAMPLE(0.1 PERCENT);

SELECT * FROM source LIMIT 10
```

Hive – Group By

```
SELECT
        stock_symbol,
        dt,
        COUNT(*)
FROM
        mytable
GROUP BY
        stock_symbol,
        dt;
```

Hive – Filter

```
SELECT * FROM mytable WHERE stock_symbol = 'NAC';

SELECT * FROM mytable WHERE stock_symbol == 'NAC';
```