MySQL and Hadoop: Big Data Integration

Shubhangi Garg & Neha Kumari
MySQL Engineering

Agenda

- Design rationale
- Implementation
- Installation
- Schema to Directory Mappings
- Integration with Hive
- Roadmap
- Q&A

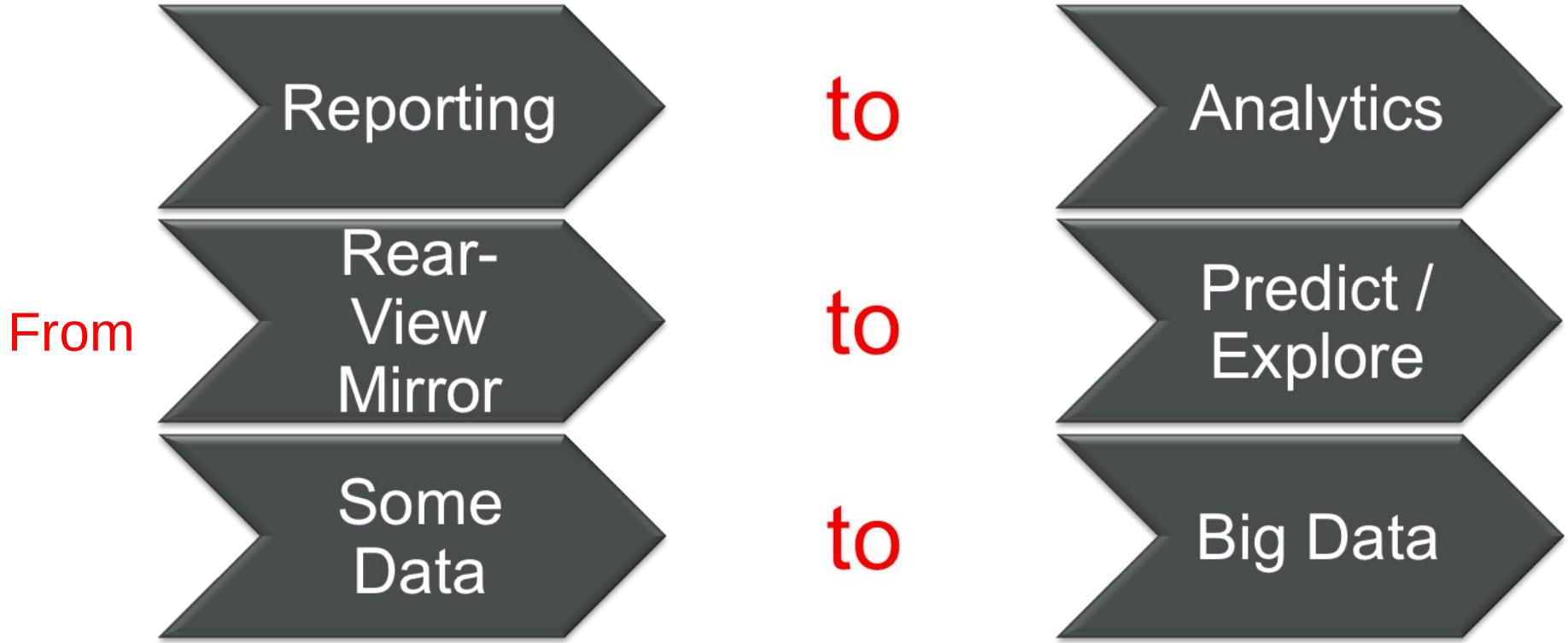
MySQL ORACLE

Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decision. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

MySQL ORACLE

Big Data: Strategic Transformation



MySQL ORACLE

CIO & Business Priority

- Web recommendations
- Sentiment Analysis
- Marketing Campaign Analysis
- Customer Churn Modeling
- Fraud Detection
- Research and Development
- Risk Modeling
- Machine Learning

FORTUNE 500

90% with Pilot Projects at end of 2012

Poor Data Costs 35% in Annual Revenues

10% Improvement in Data Usability Drives $2bn in Revenue

MySQL ORACLE

CIO & Business Priority

US HEALTH CARE	MANUFACTURING	GLOBAL PERSONAL LOCATION DATA	EUROPE PUBLIC SECTOR ADMIN	US RETAIL
Increase industry value per year by	Decrease dev., assembly costs by	Increase service provider revenue by	Increase industry value per year by	Increase net margin by
$300 B	**–50%**	**$100 B**	**€250 B**	**60+%**

"In a big data world, a competitor that fails to sufficiently develop its capabilities will be left behind."

McKinsey Global Institute

MySQL ORACLE

Analysts on Big Data

"The **area of greatest interest to my clients is Big Data** and its role in helping businesses understand customers better."

Michael Maoz, ***Gartner***

"Almost **half of IT departments in enterprises** in North America, Europe and Asia-Pacific plan to invest in Big Data analytics in the near future."

Tony Baer, ***Ovum***

"Big Data Will Help **Shape Your Market's Next Big Winners**."

Brian Hopkins, ***Forrester***

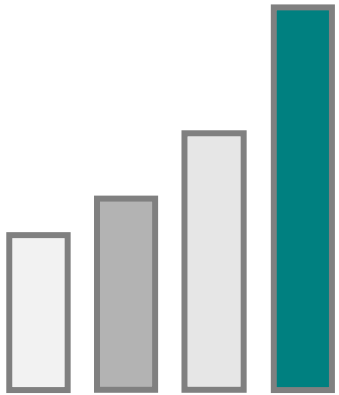
"CIOs will need to be realistic about their approach to 'Big Data' analytics and **focus on specific use cases where it will have the biggest business impact."**

Philip Carter, ***IDC***

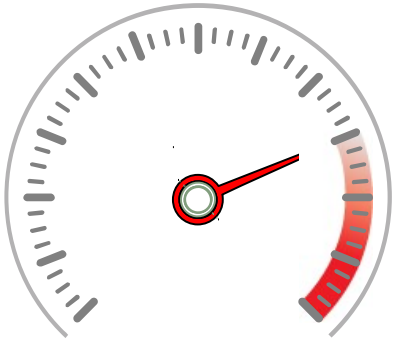
MySQL ORACLE

What Makes it Big Data?

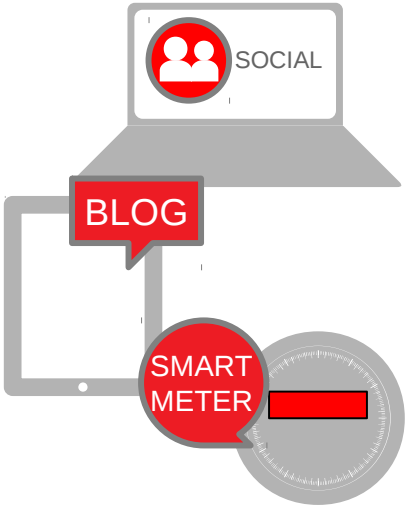
PROBLEM: Exceeds limits of conventional systems



VOLUME



VELOCITY



VARIETY

VARIABILITY

What's Changed?

- Enablers
  - Digitization – *nearly* everything has a digital heartbeat
  - Now practical to store much larger data volumes (distributed file systems)
  - Now practical to process much larger data volumes (parallel processing)
- Why is this different from BI/DW?
  - Business formulated questions to ask upfront
  - Drove what was data collected, data model, query design

Big Data Complements Traditional Methods:
Enables what-if analysis, real-time discovery

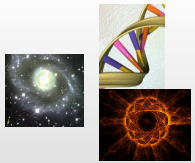
MySQL ORACLE

Unlocking Value of ALL Data

Big Data:

Decisions based on all your data

Video and Images



Documents

Social Data

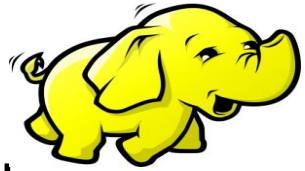
Machine-Generated Data

Traditional Architecture:

Decisions based on database data

Transactions

Why Hadoop?



- Scales to thousands of nodes, TB of structured and unstructured data
  - Combines data from multiple sources, schemaless
  - Run queries against all of the data
- Runs on commodity servers, handle storage and processing
- Data replicated, self-healing
- Initially just batch (Map/Reduce) processing
  - Extending with interactive querying, via Apache Drill, Cloudera Impala, Stinger etc.

MySQL ORACLE

MySQL + Hadoop:
Unlocking the Power of Big Data

50% of our users integrate with MySQL*

Download the MySQL Guide to Big Data:
http://www.mysql.com/why-mysql/white-papers/mysql-and-hadoop-guide-to-big-data-integration/

*Leading Hadoop Vendor

MySQL ORACLE

Leading Use-Case, On-Line Retail

Users

Browsing

Recommendations

MySQL

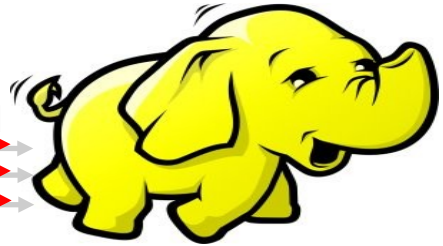
Profile,
Purchase
History

Recommendations

Social media updates
Preferences
Brands “Liked”

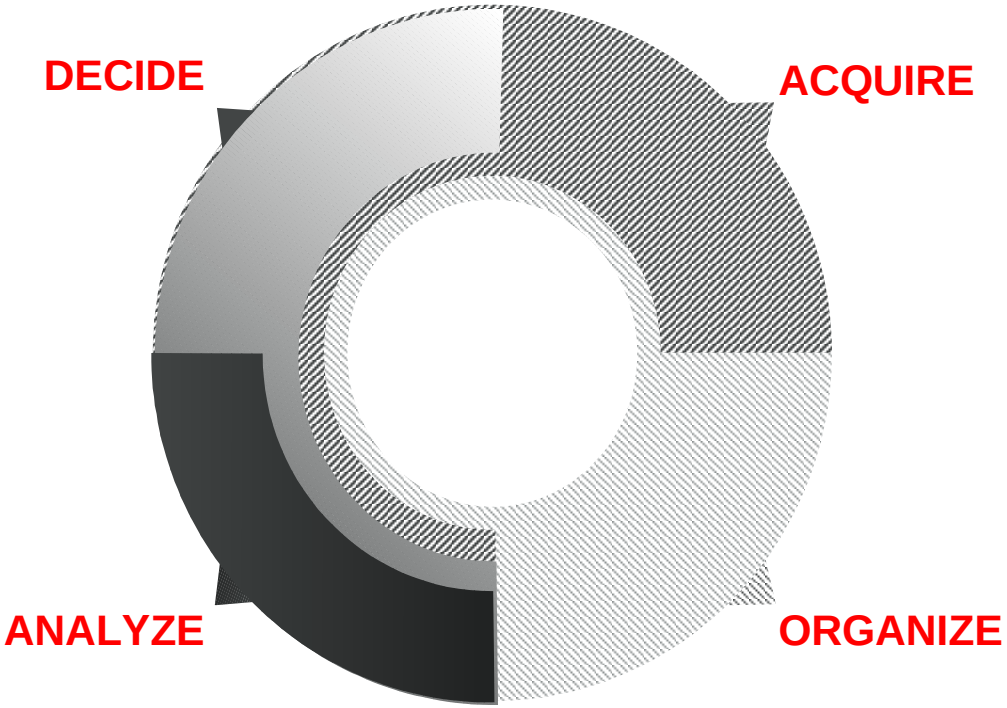
Web Logs:
Pages Viewed
Comments Posted

Telephony Stream

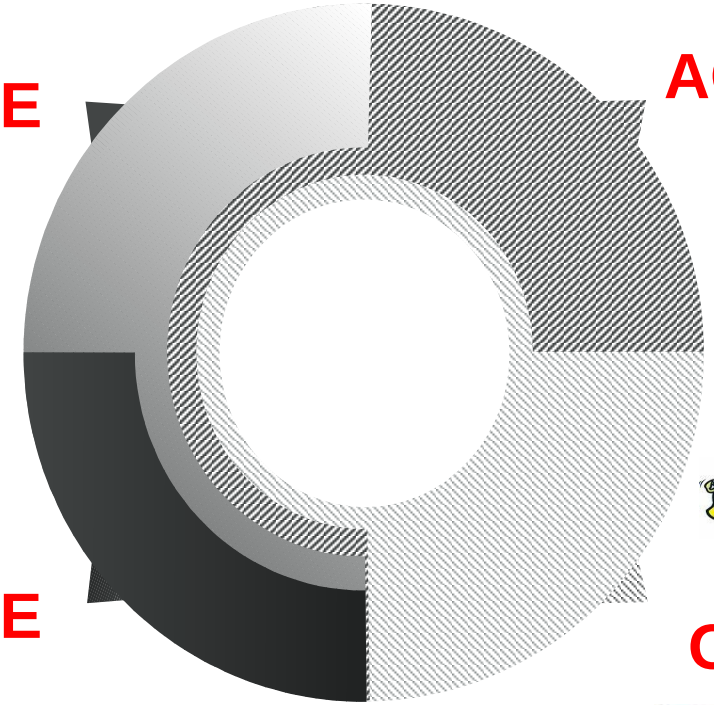


MySQL ORACLE

MySQL Applier for Hadoop: Big Data Lifecycle



MySQL in the Big Data Lifecycle



BI Solutions

MySQL

Sqoop

MySQL Applier

hadoop

hadoop

Sqoop

MySQL ORACLE

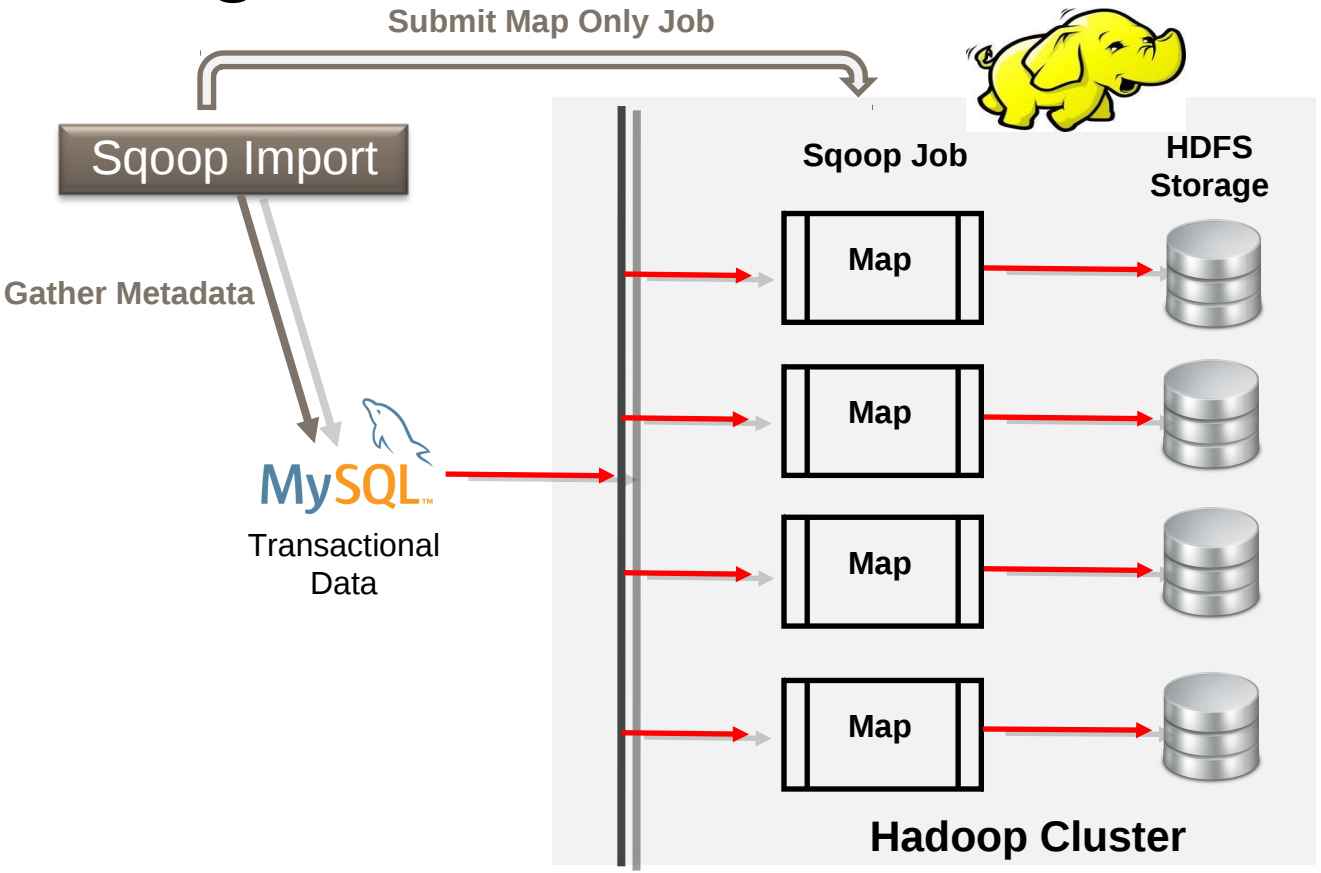
Apache Sqoop

- Apache TLP, part of Hadoop project
  - Developed by Cloudera
- Bulk data import and export
  - Between Hadoop (HDFS) and external data stores
- JDBC Connector architecture
  - Supports plug-ins for specific functionality
- “Fast Path” Connector developed for MySQL

The Apache Software Foundation
http://www.apache.org/

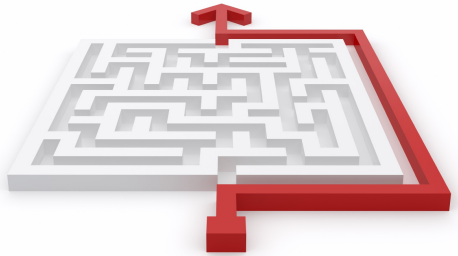
MySQL ORACLE

Importing Data

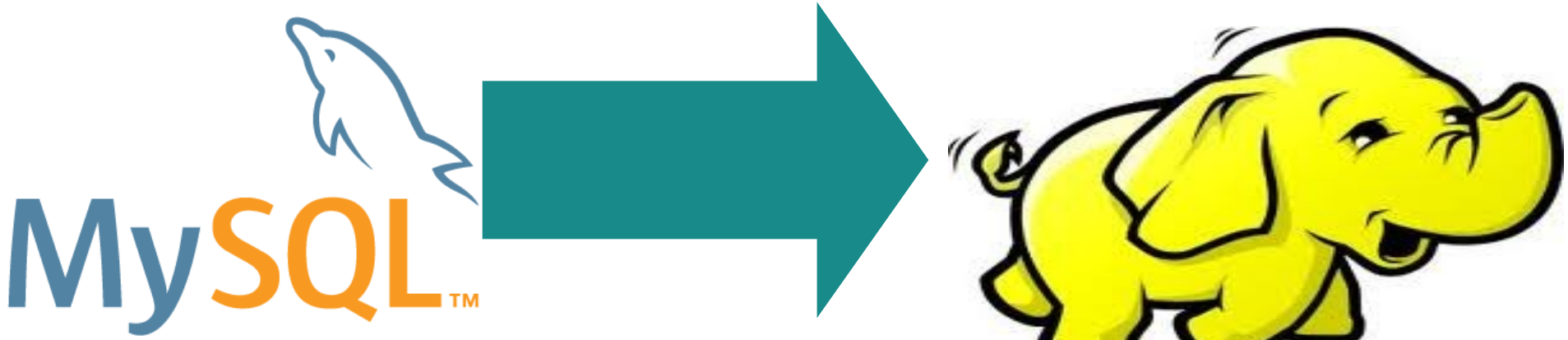


Ensure Proper Design

- **Performance impact:** bulk transfers to and from operational systems
- **Complexity:** configuration, usage, error reporting



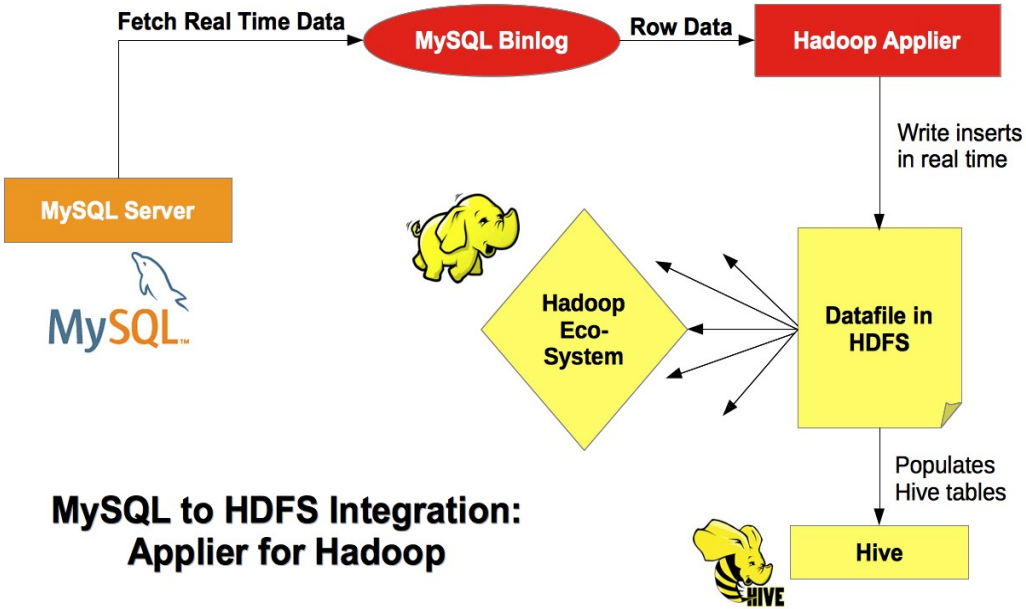
MySQL Applier for Hadoop



Real-Time Event Streaming
MySQL to HDFS

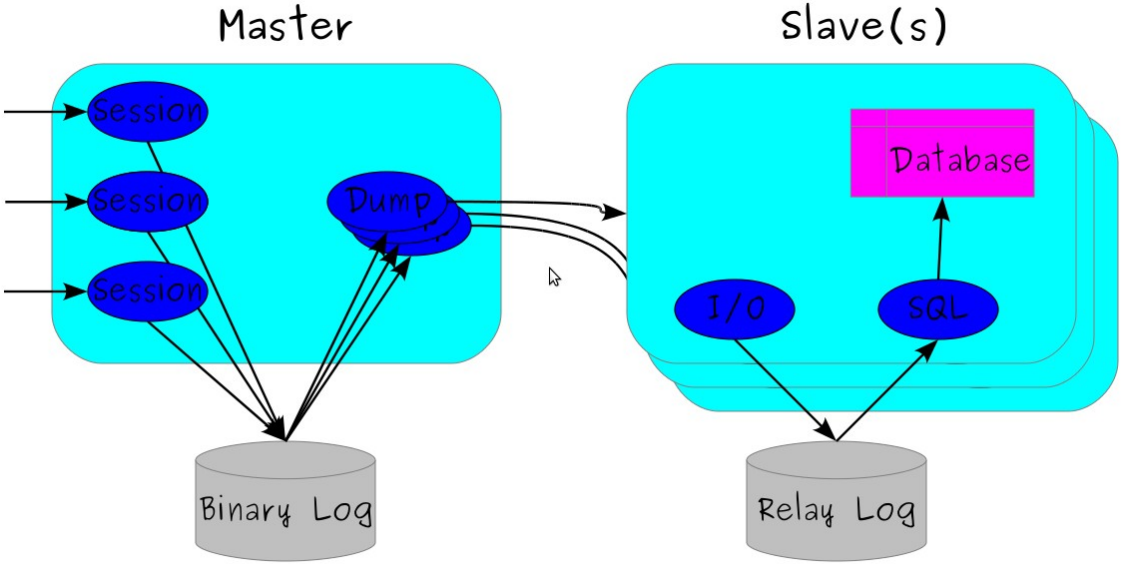
New: MySQL Applier for Hadoop

- Real-time streaming of events from MySQL to Hadoop
  - Supports move towards “Speed of Thought” analytics
- Connects to the binary log, writes events to HDFS via libhdfs library
- Each database table mapped to a Hive data warehouse directory
- Enables eco-system of Hadoop tools to integrate with MySQL data
- See dev.mysql.com for articles
- Available for download now
  - labs.mysql.com



MySQL Applier for Hadoop: Basics

- Replication Architecture

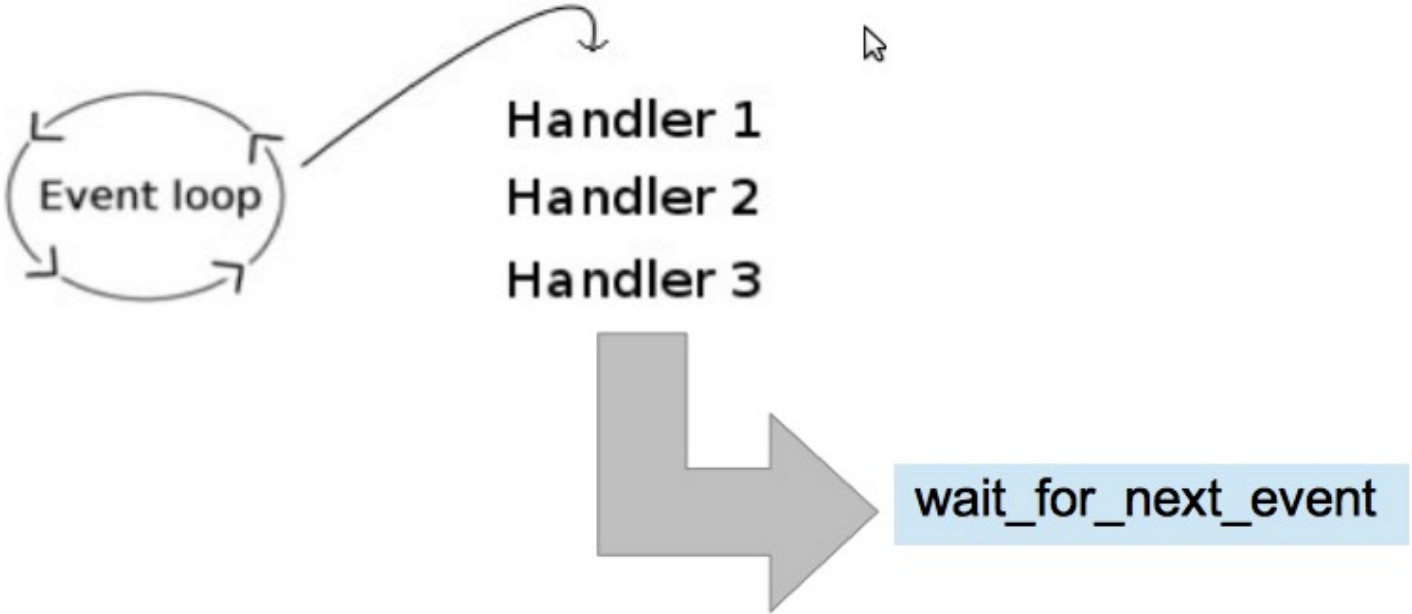


MySQL Applier for Hadoop: Basics

- What is MySQL Applier for Hadoop ?
  - An utility which will allow you to transfer data from MySQL to HDFS.
  - Reads binary log from server on a real time basis
  - Uri for connecting to HDFS: const *uri= “hdfs://user@localhost:9000”;
  - Network Transport: const *uri= “mysql://root@127.0.0.1:3306”;
- Decode binary log events
  - Contain code to decode the events
  - Uses the user defined content handler, and if nothing is specified then the default one in order to process events
  - Cannot handle all events
- Event Driven API

MySQL Applier for Hadoop: Basics

- Event driven API: Content Handlers



MySQL Applier for Hadoop: Implementation

- Replicates rows inserted into a table in MySQL to Hadoop Distributed File System
- Uses an API provided by libhdfs, a C library to manipulate files in HDFS
- The library comes pre-compiled with Hadoop Distributions
- Connects to the MySQL master (or reads the binary log generated by MySQL) to:
  - Fetch the row insert events occurring on the master
  - Decode these events, extracting data inserted into each field of the row
  - Separate the data by the desired field delimiters and row delimiters
  - Use content handlers to get it in the format required
  - Append it to a text file in HDFS

Installation: Pre-requisites

- Hadoop Applier package from http://labs.mysql.com
- Hadoop 1.0.4 or later
- Java version 6 or later (since Hadoop is written in Java)
- libhdfs (it comes pre compiled with Hadoop distros)
- Cmake 2.6 or greater
- Libmysqlclient 5.6
- Openssl
- Gcc 4.6.3
- MySQL Server 5.6
- FindHDFS.cmake (cmake file to find libhdfs library while compiling)
- FindJNI.cmake (optional, check if you already have one):

  $locate FindJNI.cmake

- Hive (optional)

Installation: Steps

- Download a copy of Hadoop Applier from http://labs.mysql.com
- Download a Hadoop release, configure dfs to set the append property true
- (**flag dfs.support.append**)
- Set the environment variable $HADOOP_HOME
- Ensure that 'FindHDFS.cmake' is in the CMAKE_MODULE_PATH
- libhdfs being JNI based, JNI header files and libraries are also required
- Build and install the library 'libreplication', using cmake
- Set the CLASSPATH

  $export PATH= $HADOOP_HOME/bin:$PATH

  $export CLASSPATH= $(hadoop classpath)
- Compile by the command “make happlier” on the terminal.
- This will create an executable file in the mysql2hdfs directory in the repository

Integration with HIVE

- Hive runs on top of Hadoop. Install HIVE on the hadoop master node
- Set the default datawarehouse directory same as the base directory into which Hadoop Applier writes
- Create similar schema's on both MySQL & Hive
- Timestamps are inserted as first field in HDFS files
- Data is stored in 'datafile1.txt' by default
- The working directory is base_dir/db_name.db/tb_name

SQL Query	Hive QL
CREATE TABLE t (i INT);	CREATE TABLE t (time_stamp INT, i INT) [ROW FORMAT DELIMITED] STORED AS TEXTFILE;

MySQL ORACLE

Mapping Between MySQL and HDFS Schema

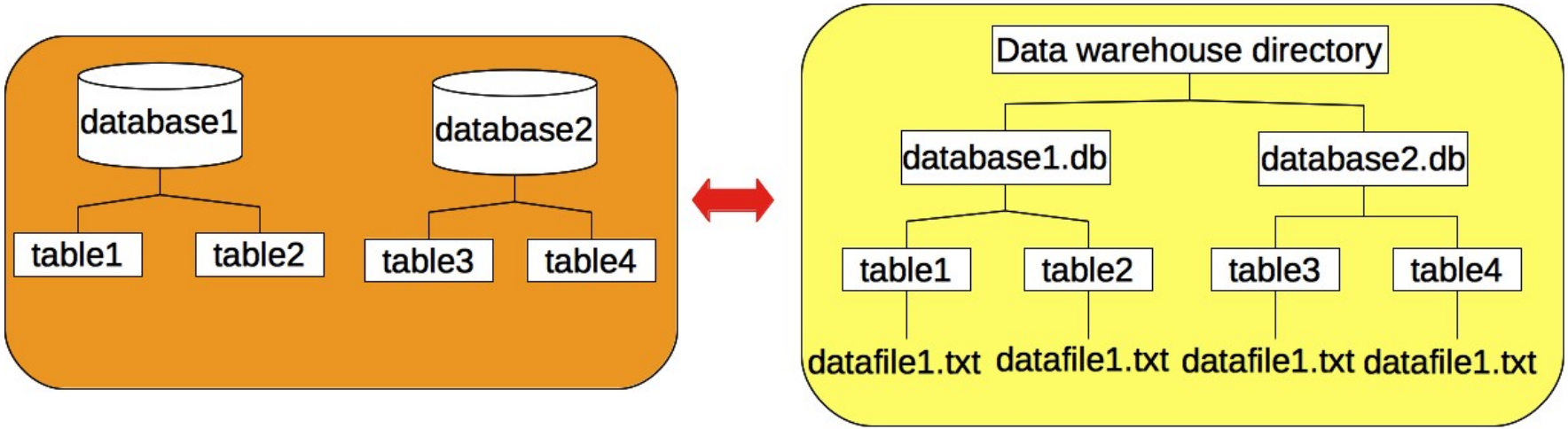


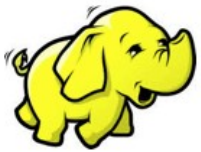
table1

col1	col2	.....
data1	data2	....
data3	data4	....

table1/datafile1.txt

ts1,data1,data2,...
ts2,data3,data4,...
....

ts=timestamp



MySQL ORACLE

Hadoop Applier in Action

- Run hadoop dfs (namenode and datanode)
- Start a MySQL server as master with row based replication
- For ex: using mtr:

$MySQL-5.6/mysql-test$./mtr --start –suite=rpl --mysqld=-binlog_format='ROW' –mysqld=--binlog_checksum=NONE

- Start hive (optional)
- Run the executable ./happlier

$./happlier [mysql-uri] [hdfs-uri]

- Data being replicated can be controlled by command line options.

$./happlier –help

Find the demo here:
http://www.youtube.com/watch?feature=player_detailpage&v=mZRAtCu3M1g

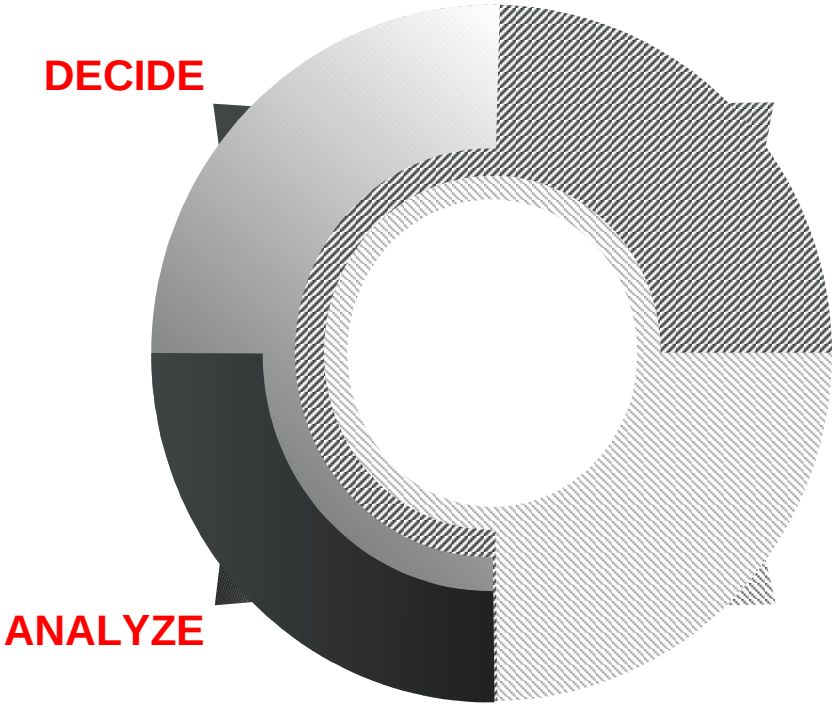
Options	Use
-r, --field-delimiter=DELIM	String separating the fields in a row
-w, --row-delimiter=DELIM	String separating the rows of a table
-d, --databases=DB_LIST Ex: d=db_name1-table1-table2,dbname2-table1,dbname3	Import entries for some databases, optionally include only specified tables.
-f, --fields=LIST	Import entries for some fields only in a table
-h, --help	Display help

Future Road Map – Under Consideration

- Support all kinds of events occurring on the MySQL master
- Support binlog checksums and GTID's, the new features in MySQL 5.6 GA
- Considering support for DDL's
- Considering support for updates and deletes
- Leave comments on the blog !

http://innovating-technology.blogspot.co.uk/2013/04/mysql-hadoop-applier-part-2.html

MySQL in the Big Data Lifecycle



Analyze
Export Data
Decide

MySQL ORACLE

Analyze Big Data

MySQL ORACLE

Summary

- MySQL + Hadoop: proven big data pipeline
- Real-Time integration: MySQL Applier for Hadoop
- Perfect complement for Hadoop interactive query tools

Integrate for Insight

Next Steps

Read the DevZone

http://dev.mysql.com/tech-resources/articles/mysql-hadoop-applier.html

Try Out Applier for Hadoop

http://labs.mysql.com

Take the Dev Quick Poll

https://dev.mysql.com/tech-resources/quickpolls/

Thank you!