

[www.qconferences.com](http://www.qconferences.com)  
[www.qconbeijing.com](http://www.qconbeijing.com)



伦敦 | 北京 | 东京 | 纽约 | 圣保罗 | 上海 | 旧金山

London · Beijing · Tokyo · New York · Sao Paulo · Shanghai · San Francisco

# QCon全球软件开发大会

International Software Development Conference

# Building Next-Generation Data Integration Platform

George Xiong  
eBay Data Platform Architect  
April 21, 2013

>50 TB/day new data

100+ Subject Areas

>100 Trillion pairs of information

>100 PB/day

Processed

>60k chains of logic

>7500

business users & analysts

24x7x365

Always online

>1000 Data Source

5000+ Target Data

second

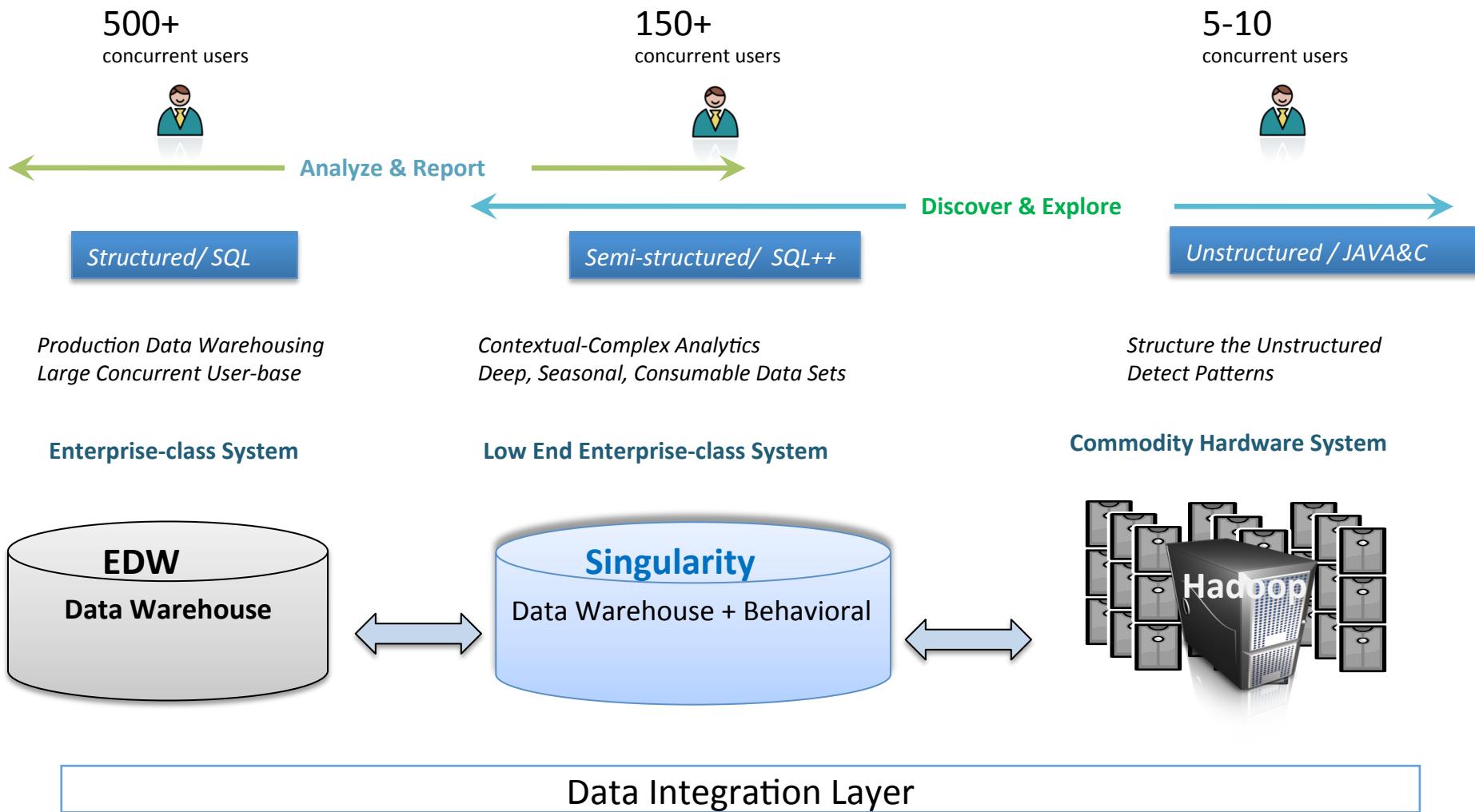
Millions of queries/day

99.98+% Availability

YEAR

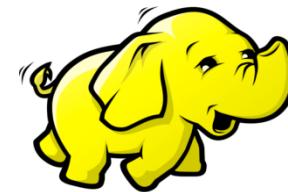
2012

# Data Platforms



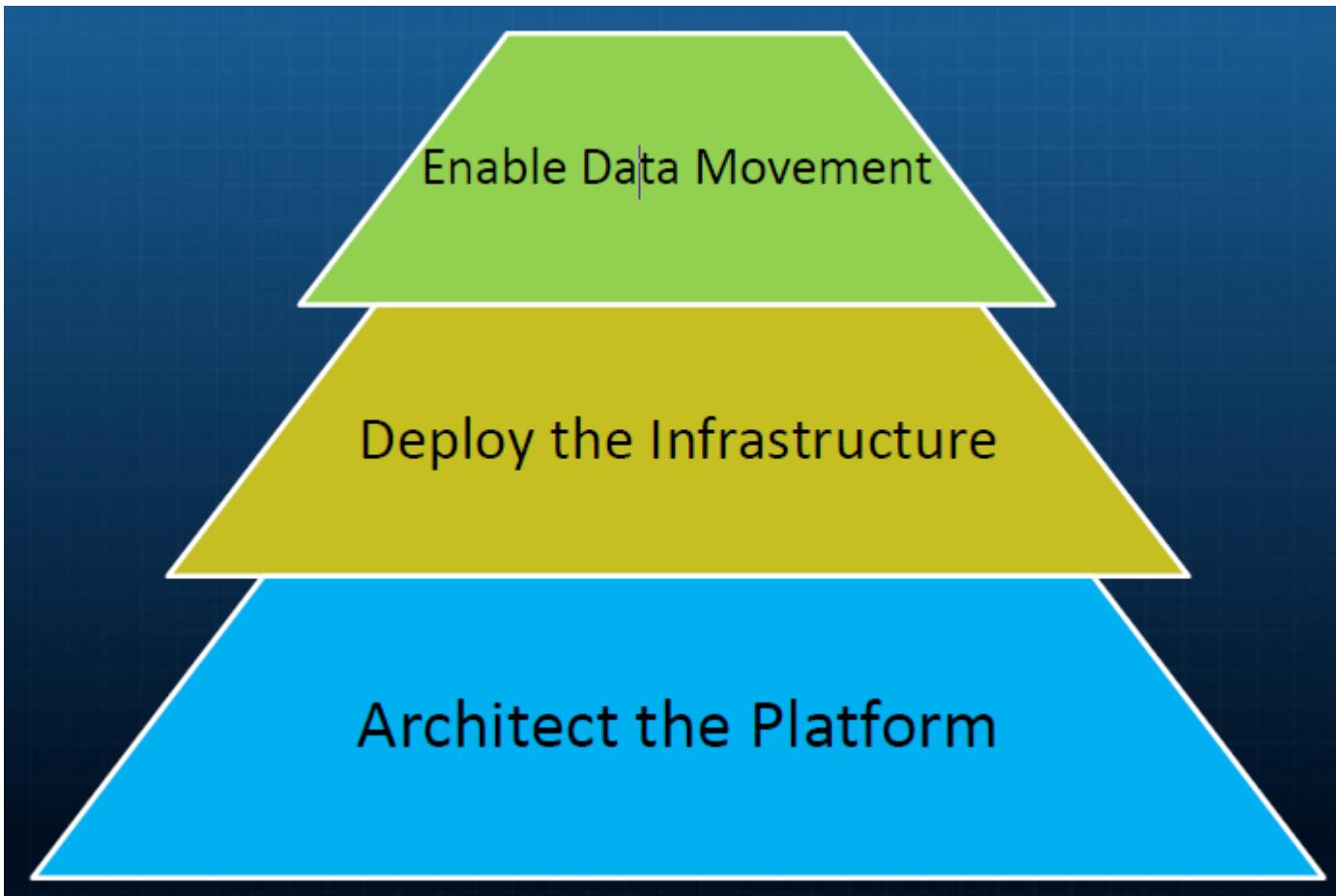
# Retrospective

- Big Data = Big Systems <> Accurate Data
- Job Complexity
- System Outage / Availability
- High Maintenance Costs
- Quick Delivery Pressure



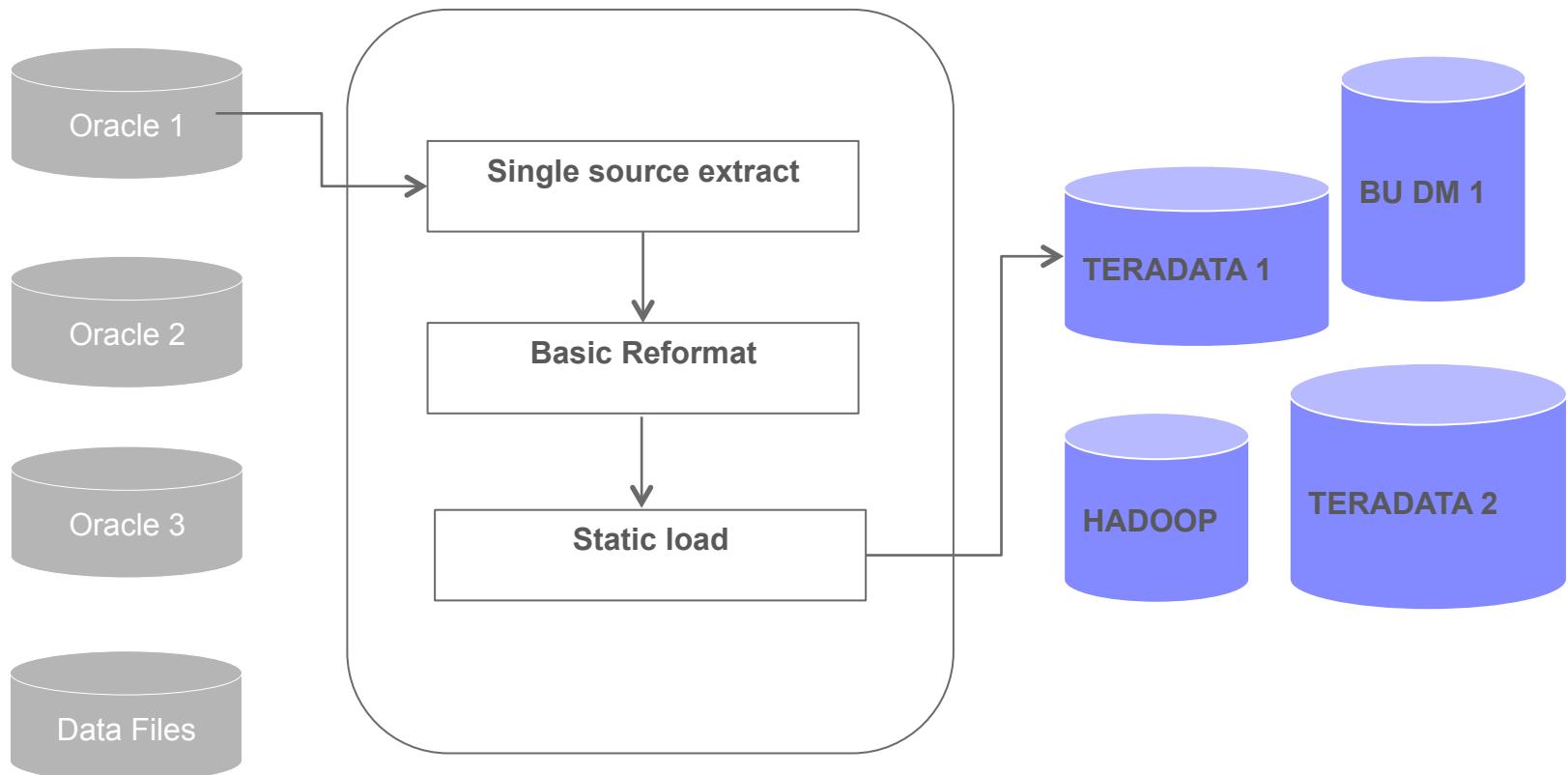
# ETL always is the first priority of DI

---



# Once upon a time

Inefficient ... Inconsistent ... In parallel  
User unfriendly



# Next-Gen ETL Requirement

Compression

Conditional Components

Multi-source/Multi-Target

Abstraction

Platform Cost Efficiently

Rapid Development

Build-in HA/DR

Hyper Reusability

High Scalability

Single Version



# Building The Foundation

Reusable, metadata driven processes

Picking the right tool

Think big, implement small, increment later

Focus on efficiency where it matters

Single Version Utilities



# Abstraction: Metadata Drives Everything



## Key Component-DML

```
record
    decimal(13) id; /* DECIMAL(12) NOT NULL*/
    string(2) code; /* CHAR(2) NOT NULL*/
    string(2) iso_country; /* CHAR(2) NOT NULL*/
    string(1) summertime_ends_first = NULL; /* CHAR(1)*/
    decimal(10) summertime_ends_month = NULL; /* DECIMAL(9)*/
    decimal(10) default_currency_id = NULL; /* DECIMAL(9)*/
    decimal(10) name_res_id = NULL; /* DECIMAL(9)*/
end
```

# Environment Setup



Common setup script

ETL Process Specific Configuration

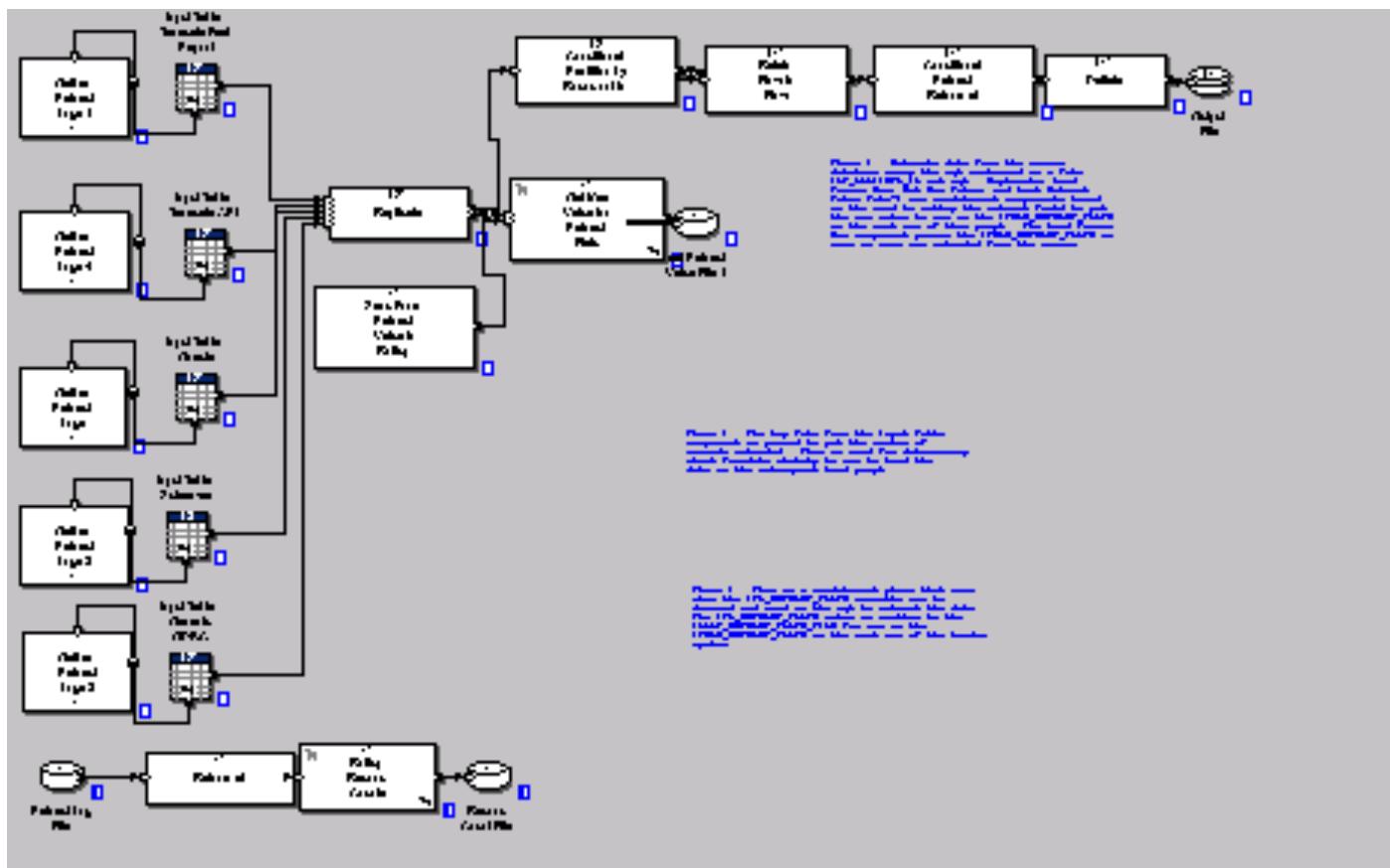
Everything evaluated at run time

## The Extract Process



- Single common extract handler
- ETL ID specific State files
- Run time metadata
- Single Module extract utility

# AB Initio Extract Graph

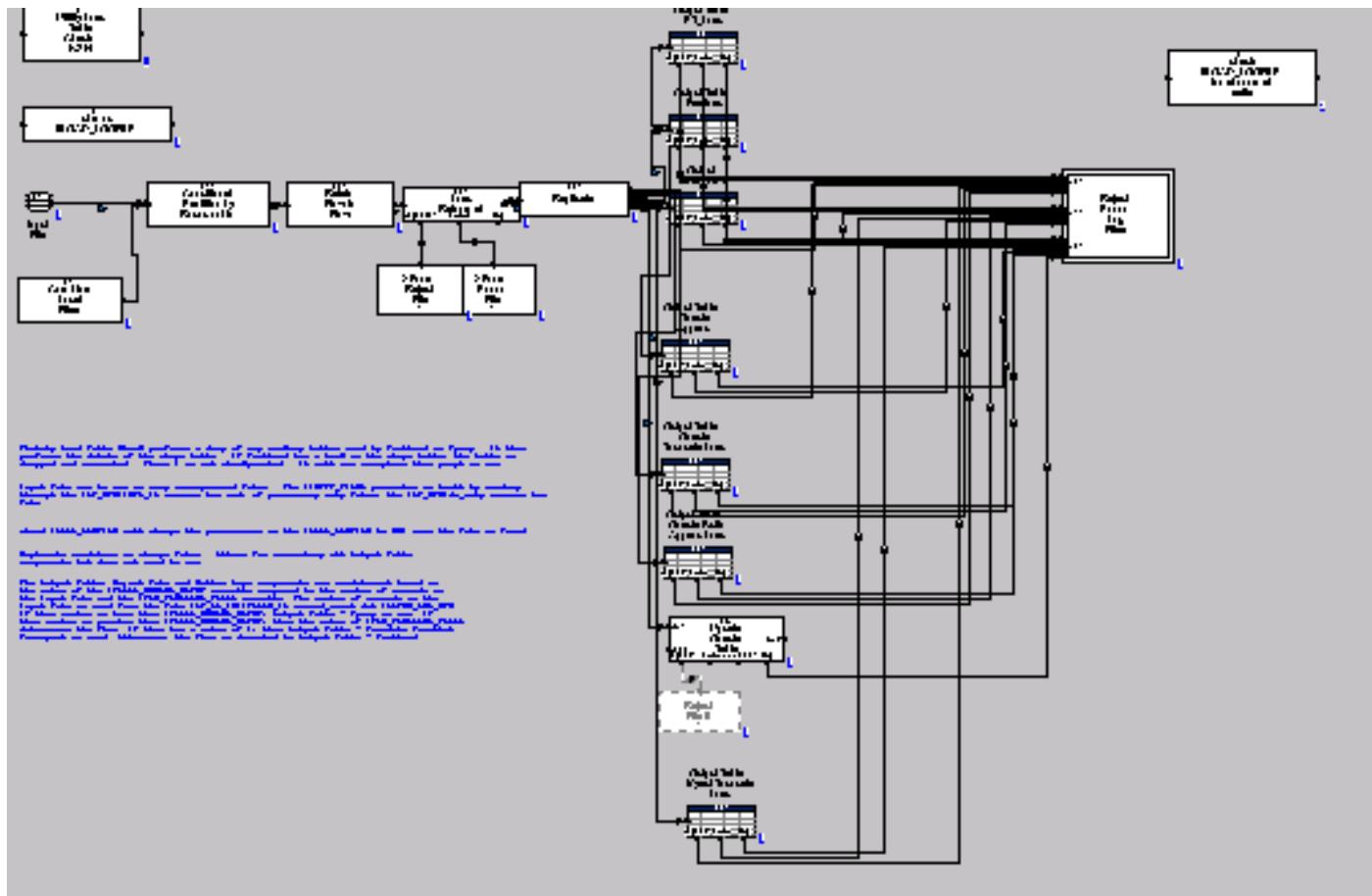


# The Load Process



- Single common Load handler
- ETL ID specific State files
- Run time metadata
- Single Module load utility
- Multi-Data Target

# AB Initio Load Graph



# The Transformation Process



- Typical Run post Load
- Dynamic environment
- Independent SQL or Mapreduce
- Run time Query Band
- Native Integrated

# The ETL Metadata



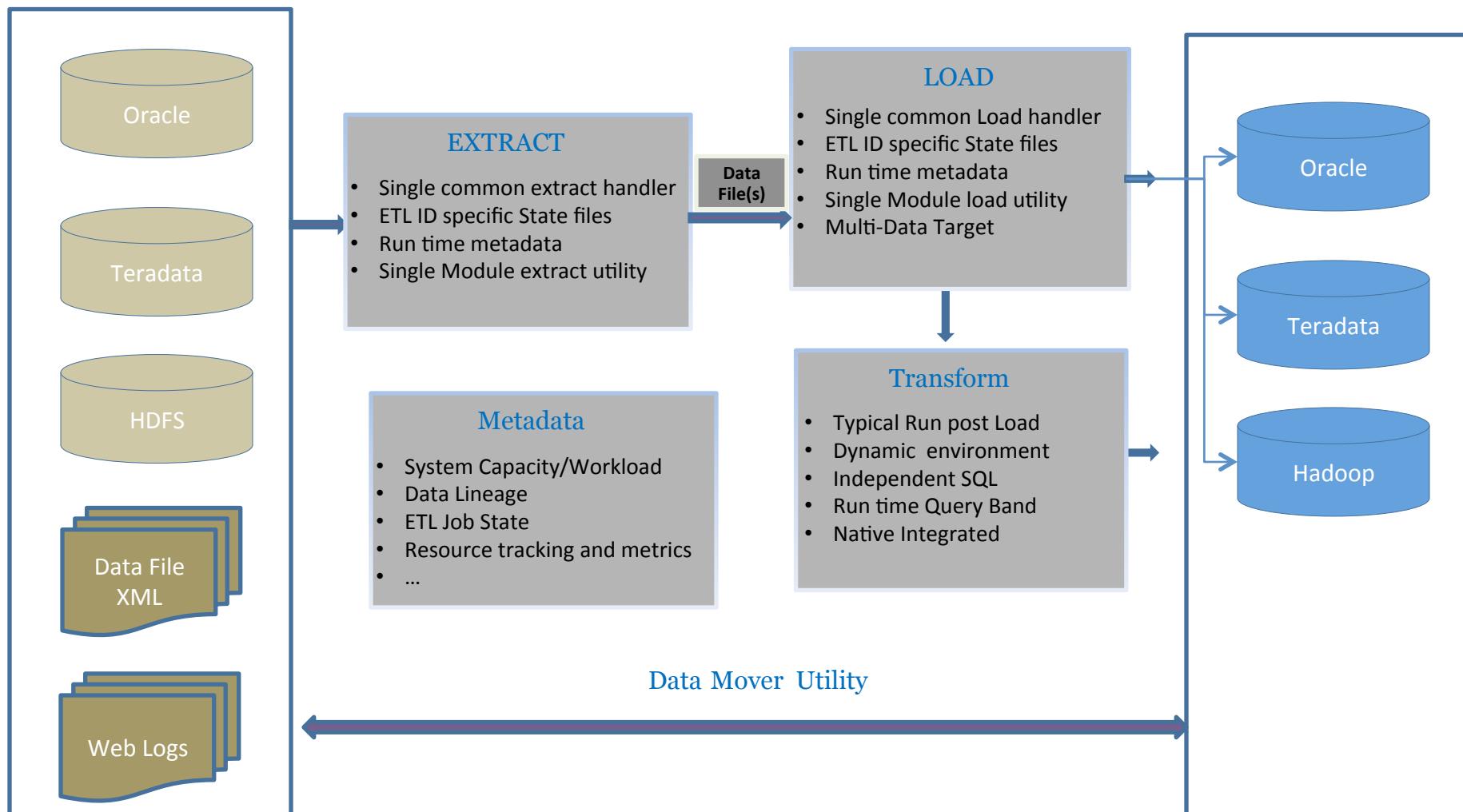
- System Capacity/Workload
- Data Lineage
- ETL Job State
- Resource tracking and metrics

## Other ETL Framework Modules



- Data Move utilities
- Unit of Work
- Data Pipeline
- ETL host Workload balance
- Job Auto Switch
- Auto ETL code smart gen tools
- ELT-> ETL
- ...

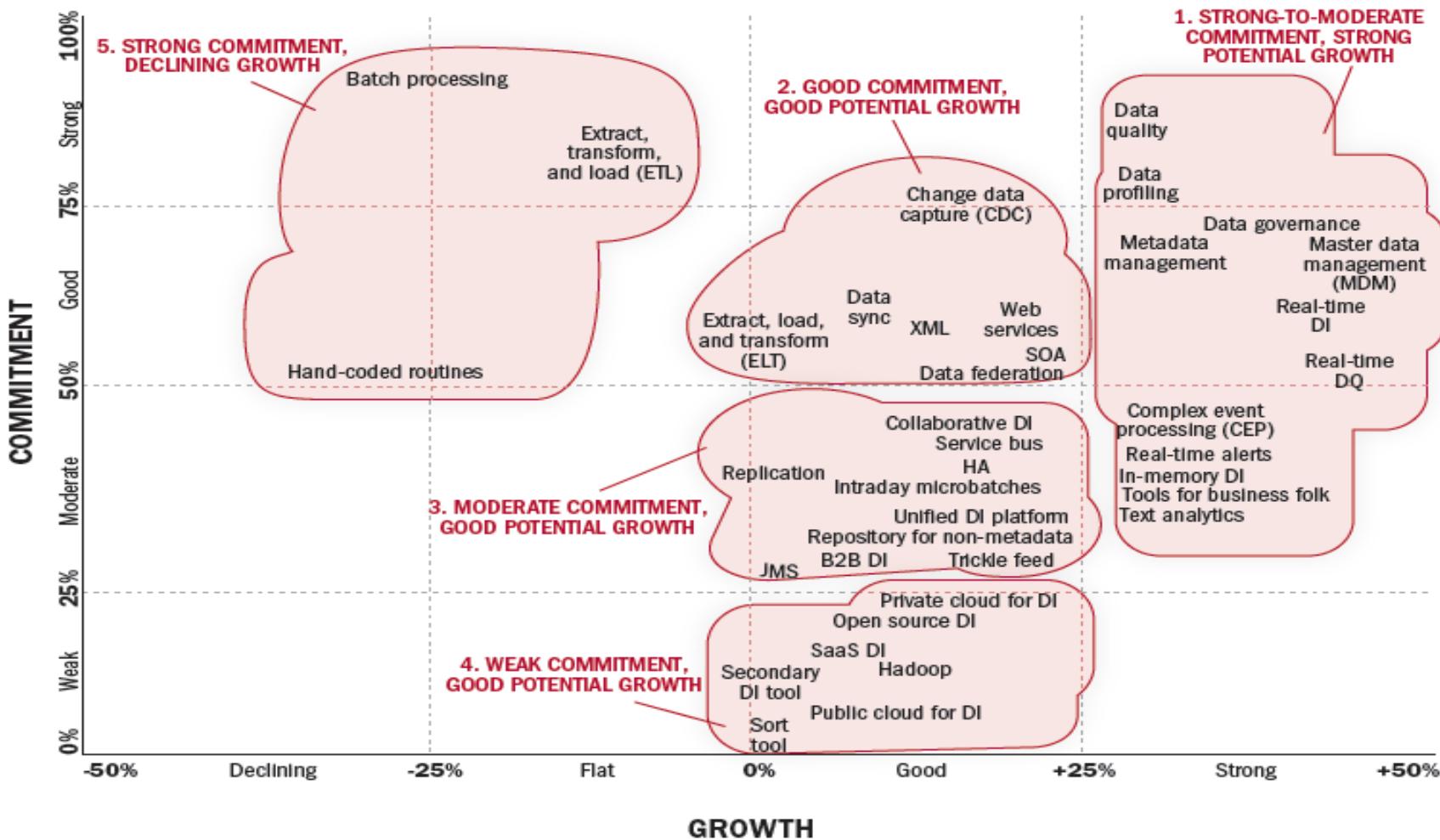
# Put It all Together



Efficient ...Consistent ...Configurable... Extensible... Parallel  
Reusability...Restart ability

# DI technologies: Not Only ETL

Next Generation DI Options Plotted for Growth and Commitment, from TDWI



# Software-as-a-service (SaaS)



>85% of eBay analytical workload is **NEW & UNKNOWN**

The metrics you know are cheap

The metrics you don't know are expensive – but high in potential ROI

**Exploration & Testing** are core pillars of an analytics-driven organization

## What is a VDM?



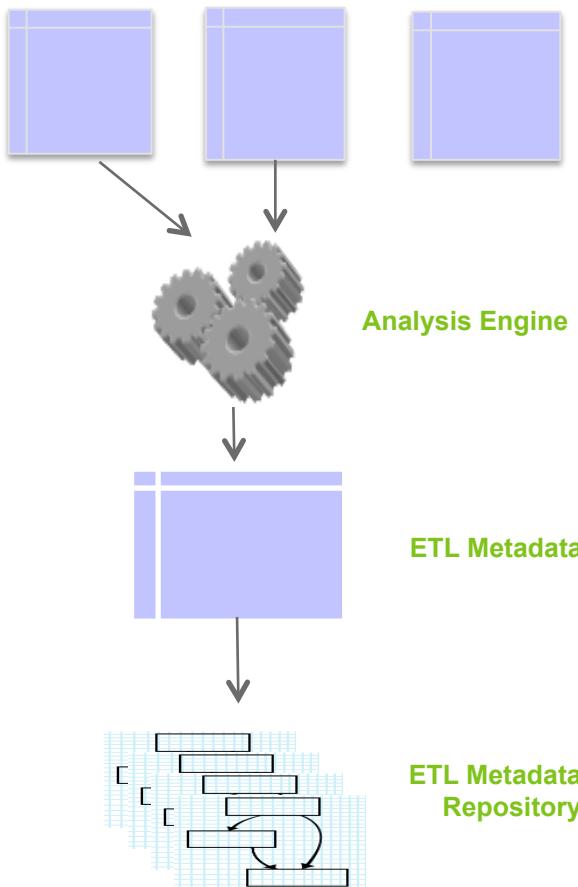
A Virtual Data Mart (VDM) is a Prototyping or Subject Area Specific Environment in Teradata (formerly called PET).

Allows End Users to create a working, non-production environment for:

- One-Time Analytics
- Specific, unique data analysis
- Loading and correlating of data from sources not currently available in the EDW
- Business Unit specific reporting and analysis

# Metadata Collecting Automation

DBQL    Table Usage Info    ETL JOB Log



**DBQL/Table Usage Info/ETL JOB LOG** are Teradata Dictionary Tables

- DBQL: Contains each query details, such as runtime, CPU cost, query band etc.
- Table Usage Info: What table(s) is been used by the query
- ETL JOB TRACKER

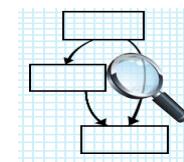
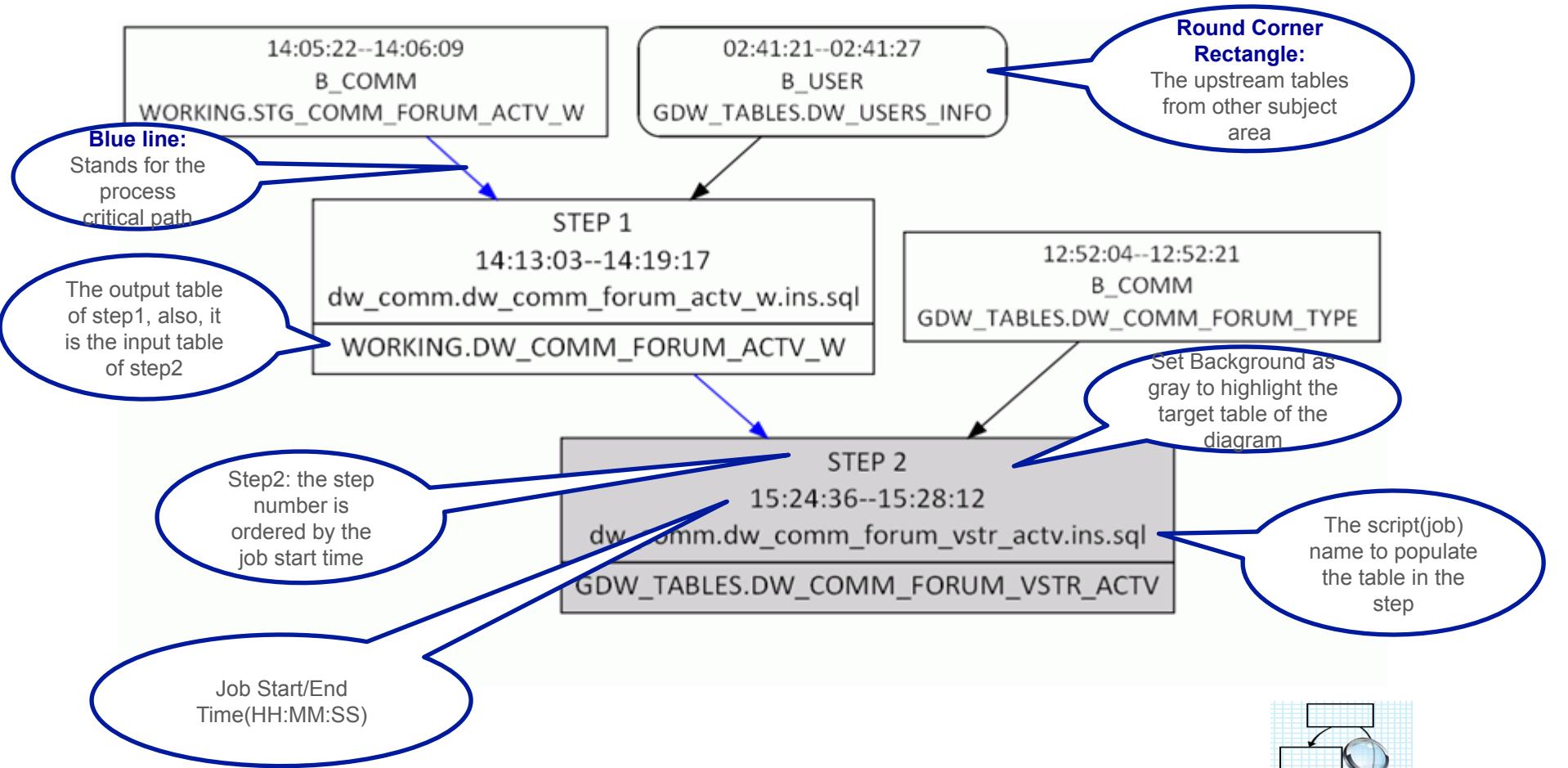
**Analysis Engine** analyze the raw data of DBQL and Table Usage Info, get dependency metadata about table(s)

- On batch script (job)level, what table(s) is output table of the script(job)
- What table(s) is input table of script(job)

**ETL Metadata** contains the result of **Analysis Engine**, including

- DFD dependency meta data of each table, with the meta data, we could draw DFD for any table via the tool Graphviz.
- Each script(job) is a node of the diagram
- The dependency between script(job) setup the mapping between nodes.

# Data Lineage



# More Data Integration Programs

Data Quality

Data Rationalization

Standardized ETL Building Tools

The Datahub

...



## Questions?

---



For More Information: [jxiong@ebay.com](mailto:jxiong@ebay.com)



@InfoQ



infoqchina

软件  
正在改变世界!