Extreme Analytics at eBay

Tom Fastner

HPTS
10/25/2011

| [ ] | All Categories ▲▼ | Search | Advanced |

- All Categories ›
- Fashion ›
- Motors ›
- Electronics ›
- Collectibles & Art ›
- Home, Outdoors & Decor ›
- Entertainment ›
- Deals & Gifts ›
- Brands on eBay
- Classifieds

Enjoy an extra
40% off storewide

Free shipping!

lastcall
BY NEIMAN MARCUS

At the new FASHION OUTLET

Ends Oct 1 at 6am PT   Shop now ›

Discount reflected in pricing.

Your last viewed items

STAR WARS FAN FAVORITES I PC DVD
$19.99
See suggestions | Remove all

Your recent searches

- tron posters
- tron memrobillia
- tron memorbellia
- tron poster
- tron costume
- tron outfits
- tron flynn

Remove searches

Sign in

Back for more fun? Sign in now to buy, bid and sell, or to manage your account.

Sign in

Not registered yet?

Join the millions of people who are already a part of the eBay family.

Register

[Hide activities]

Recommendations for you

Report an Issue (Internal only)

eBay Analytics

>50 TB/day new data

>100k data elements

>100 Trillion pairs of information

>100 PB/day Processed

>50k chains of logic

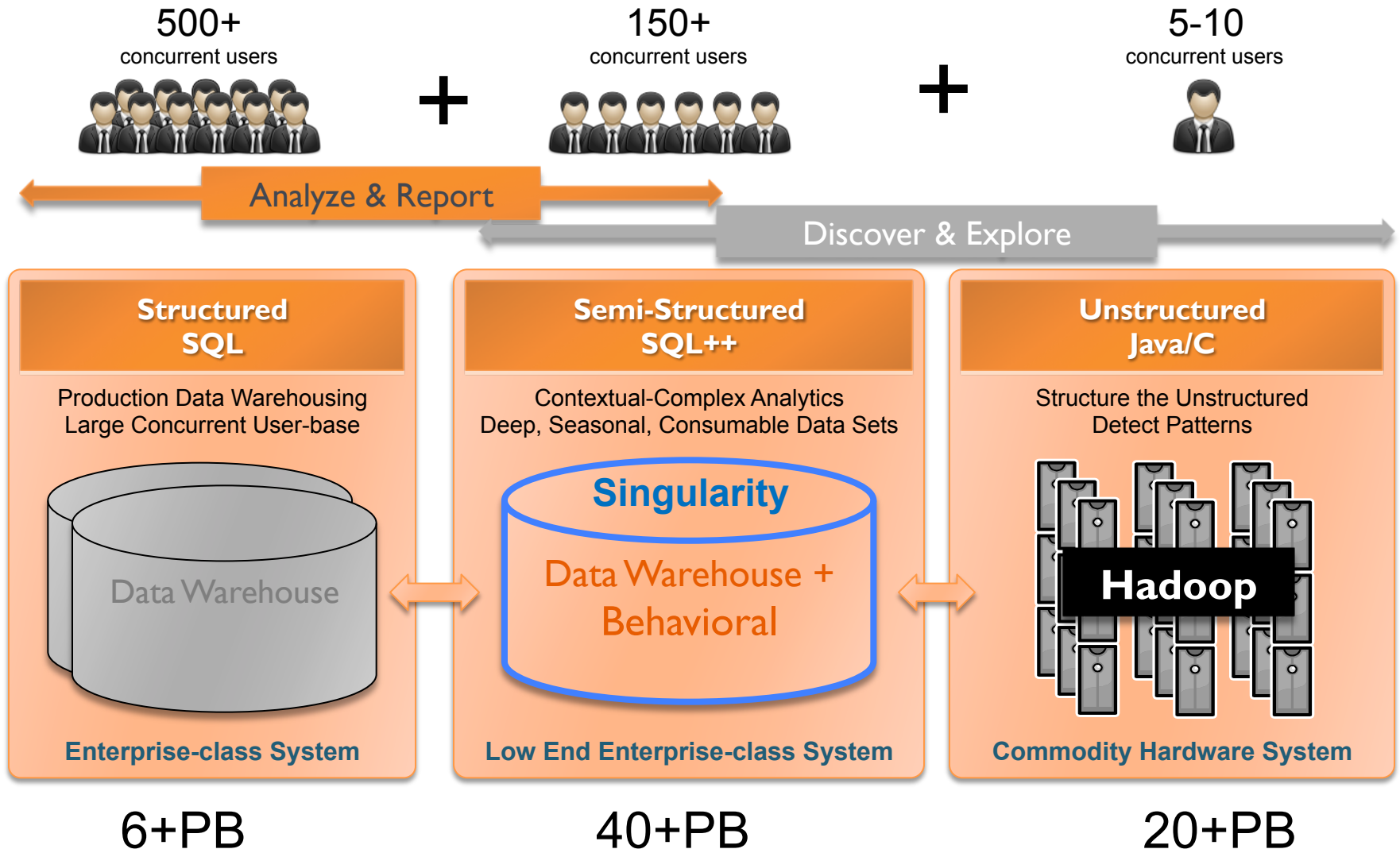>7500 business users & analysts

Active/Active

turning over a TB every second

24x7x365 Always online

Millions of queries/day

Near-Real-time

Data Platforms

500+
concurrent users

+

150+
concurrent users

+

5-10
concurrent users

Analyze & Report

Discover & Explore

Structured SQL	**Semi-Structured SQL++**	**Unstructured Java/C**
Production Data Warehousing Large Concurrent User-base	Contextual-Complex Analytics Deep, Seasonal, Consumable Data Sets	Structure the Unstructured Detect Patterns
Data Warehouse	**Singularity** Data Warehouse + Behavioral	**Hadoop**
Enterprise-class System	**Low End Enterprise-class System**	**Commodity Hardware System**

6+PB

40+PB

20+PB

ebaY
Analytics
Platform
& Delivery

Behavioral Data Flow

eBay Visitors → Application Server → CAL → Analytics Platform & Delivery → Analysts
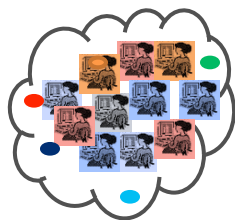
Application servers
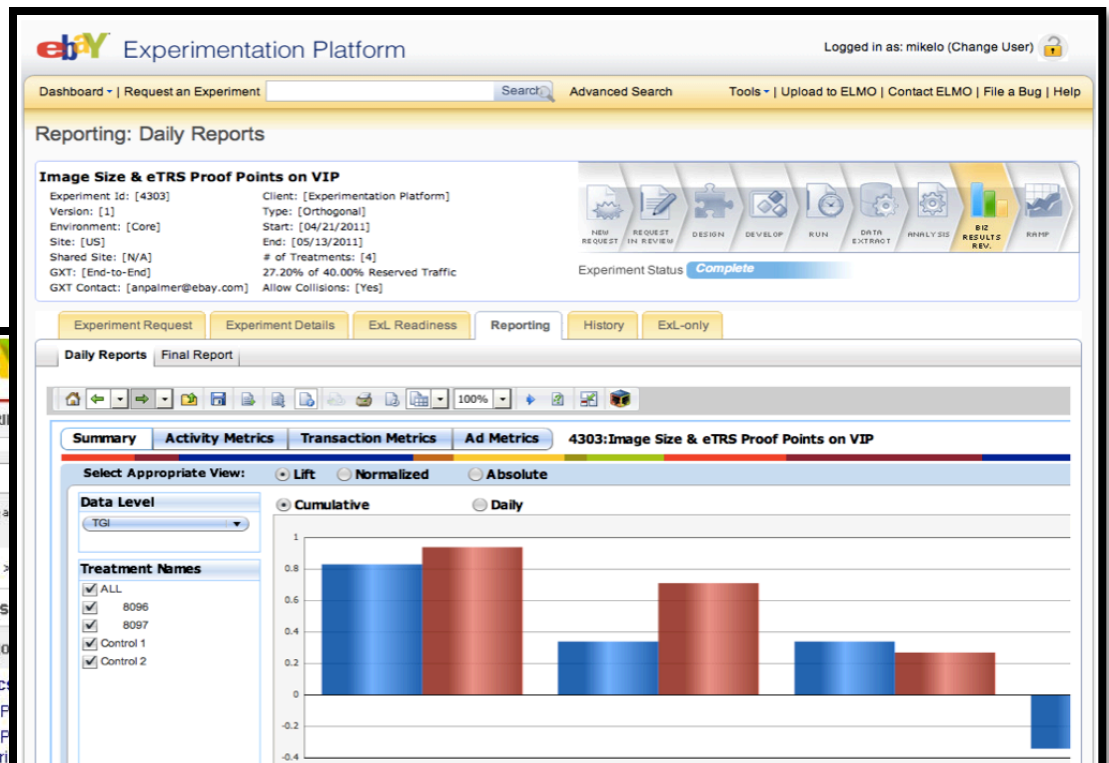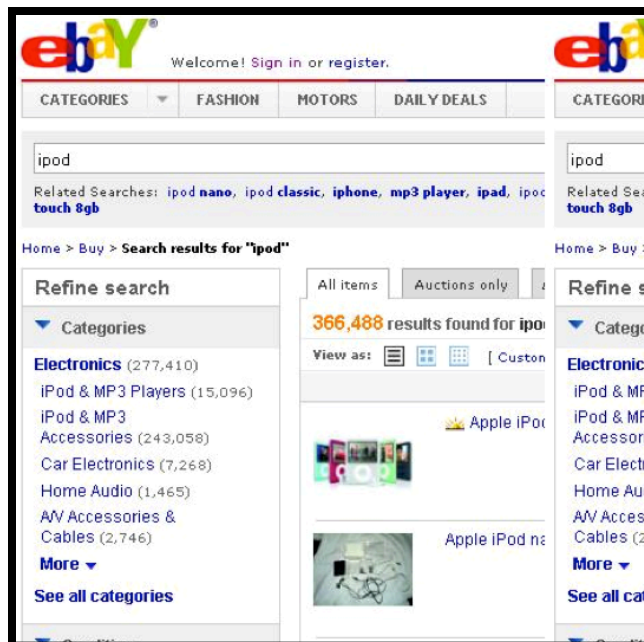
Central Application Logging (Tibco)

Listeners　　**Ingest Servers**　　**DB**　　**Singularity**　　**Hadoop**

Singularity Use Case for Semi-Structured Data

Semi-Structured Data

Start_dt	Guid	Sess_id	Page_id	Soj
2011-10-18	1234	1	15	Language=en& source=hp& itm=i1,i2,i3,i4,i5

```
SELECT start_dt, guid, sess_id, page_id,
        NVL(e.soj, 'itm') AS item_list
FROM    event e
WHERE e.start_dt = '2011-10-18'
        AND e.page_id = 15            /* Search Results */
```

Start_dt	Guid	Sess_id	Page_id	Item_list
2011-10-18	1234	1	15	i1,i2,i3,i4,i5

Semi-Structured Data

```
WITH event (start_dt, item_list) AS (<previous SQL>)
SELECT
        start_dt,
        item_id,         /* Individual Item */
        count(*)
FROM    TABLE (              /* Normalize comma delimited list */
                normalize_list( start_dt, item_list, ',')
                RETURNS(start_dt, idx, item_id)
        )
GROUP BY 1, 2
ORDER BY 3 DESC
```

*syntax simplified

Start_dt	Item_id	Count(*)
2011-10-18	i1	555
2011-10-18	i2	444
2011-10-18	i3	333
2011-10-18	i4	222
2011-10-18	i5	111

Semi-Structured Data

Event Table

~ 4 Billion rows per Day

~ 2 Trillion rows in ~640 partitions (day)

~ 10,000 Tags

~ 40 Billion Search impressions per day

~ 1.2 PB compressed database space

~ 6 PB raw, uncompressed data

32 Seconds

Start_dt	Item_id	Count(*)
2011-10-18		~70,000
2011-10-18	~135 Million Items	???
2011-10-18		???
2011-10-18		???
2011-10-18		???

Benefits of Semi-Structure

- Data Modeling
  - > No need to build out a complex model
  - > Less vulnerable to changes in the model
- ETL
  - > Simplified coding; less maintenance
  - > Less vulnerable to changes
  - > Less transformations
  - > Faster loads
  - > Better compression
- Processing
  - > Eliminating joins (NVP = pre-joined!)
  - > Having the context (rows) of "denormalized table" in one row available for processing
  - > Potentially reduce the number of path over the data using a UDF vs. like Ordered Analytics

Path Analysis

Path Analysis

Home	Search	View Item	View Item	Watch Item	Bid

H	S	V	V		W		B	
H	S		V			W		B
H	S		V (2)	W			B	

Home	MyeBay	Bid	Bid

H	M	B	B
H	M		B
H	M	B (2)	
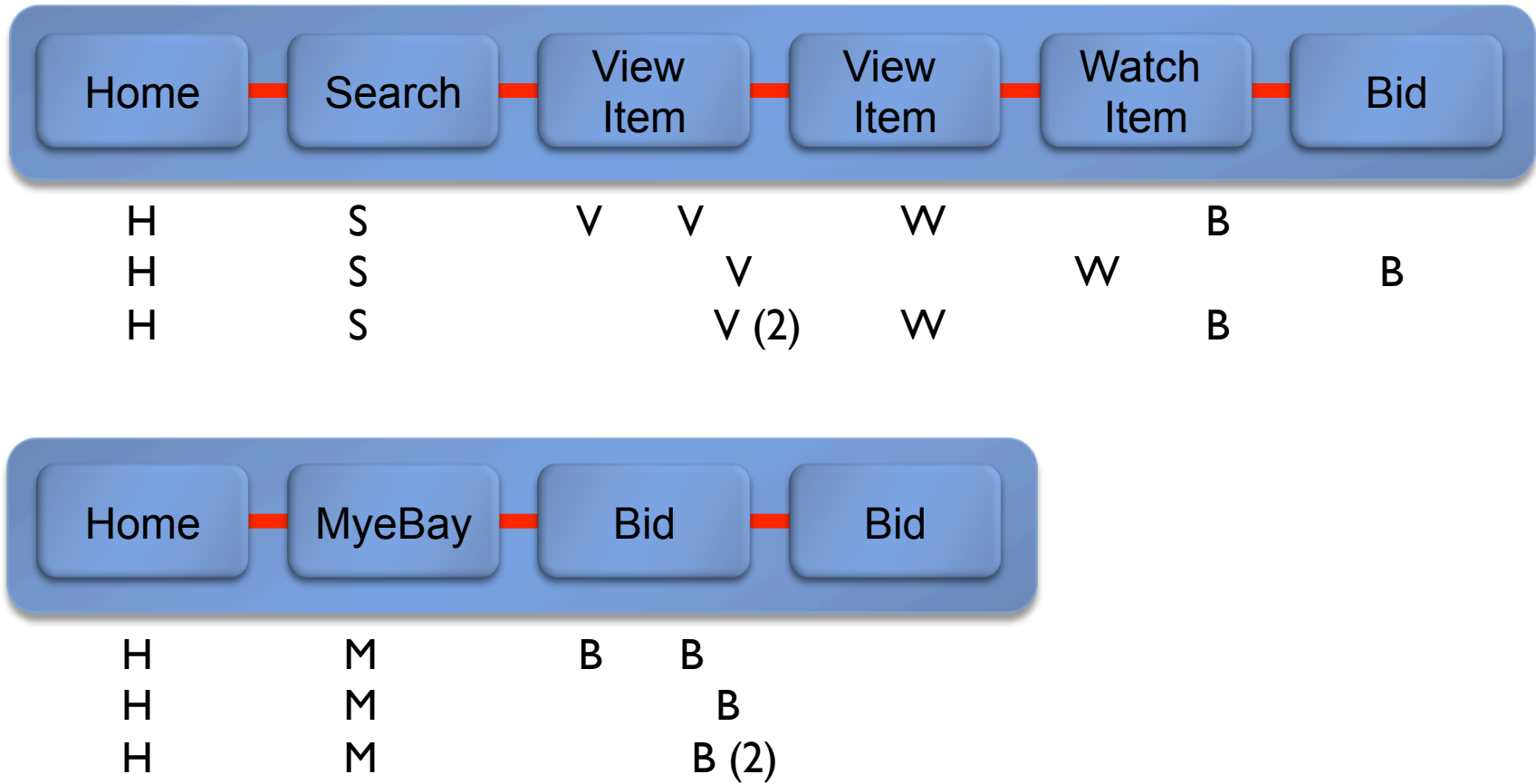
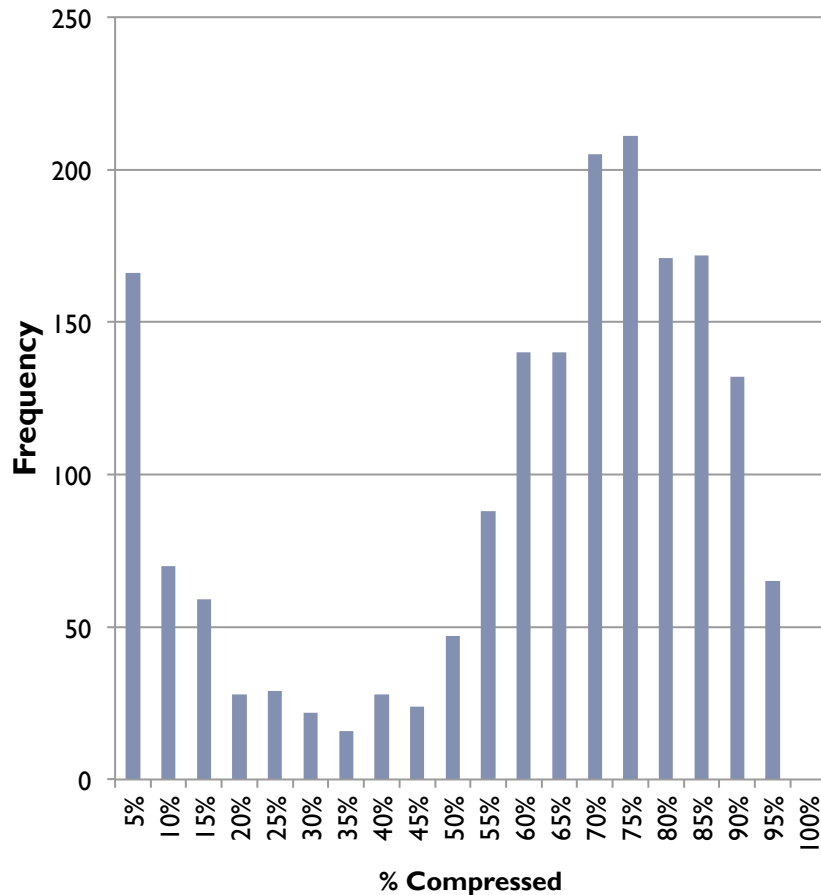# xPath Function

```
TABLE (xpath(
    new variant_type ( date, sessionid )          /* key columns */
  , 'H*B'                                          /* regex pattern */
  , 'H: pageid=1,2,3,4&B: pageid=5,6'              /* symbol definitions */
  , new variant_type ( pageid )                    /* metric columns */
                                                   /* metric definitions */
  , 'list_count(pageid of *), list_collapse(pageid of *)'
               )
    RETURNS( date, sessionid                       /* key columns */
                                                   /* metric definition columns */
  , 'm1=<count>&m2=<collapsed list>')
         )
```
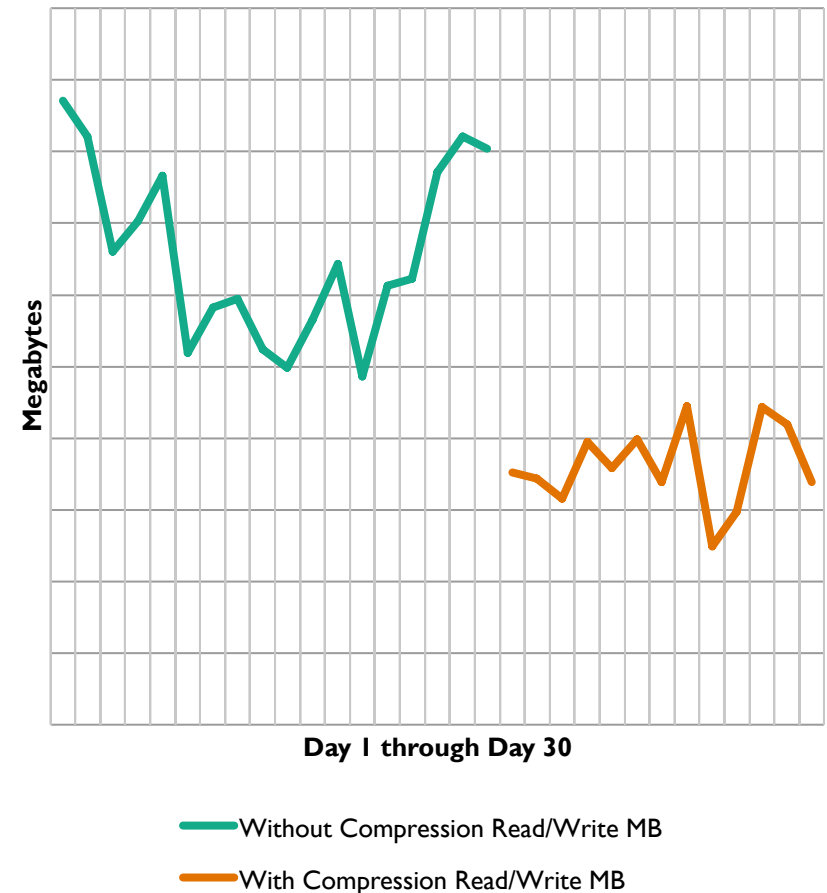
Date	Sessionid	Metric definitions
2011-10-02	1234	m1=HSV(2)WB&m2=HSVVWB
2011-10-02	5678	m1=HMB(2)&m2=HMBB

Compression (Block Level)

Compression Rate Histogram

IO Traffic by Day

Day 1 through Day 30

— Without Compression Read/Write MB

— With Compression Read/Write MB

Topology Q1/2012

Teradata-Hadoop Bridge

Teradata

SELECT Userid, Sessionid,Fraudscore
FROM TABLE (ExecMR('Athena',
'getfraudcandidates '2011-07-07');)

SELECT key, value
FROM mytable

Hadoop

getfraudcandidates '2011-07-07'
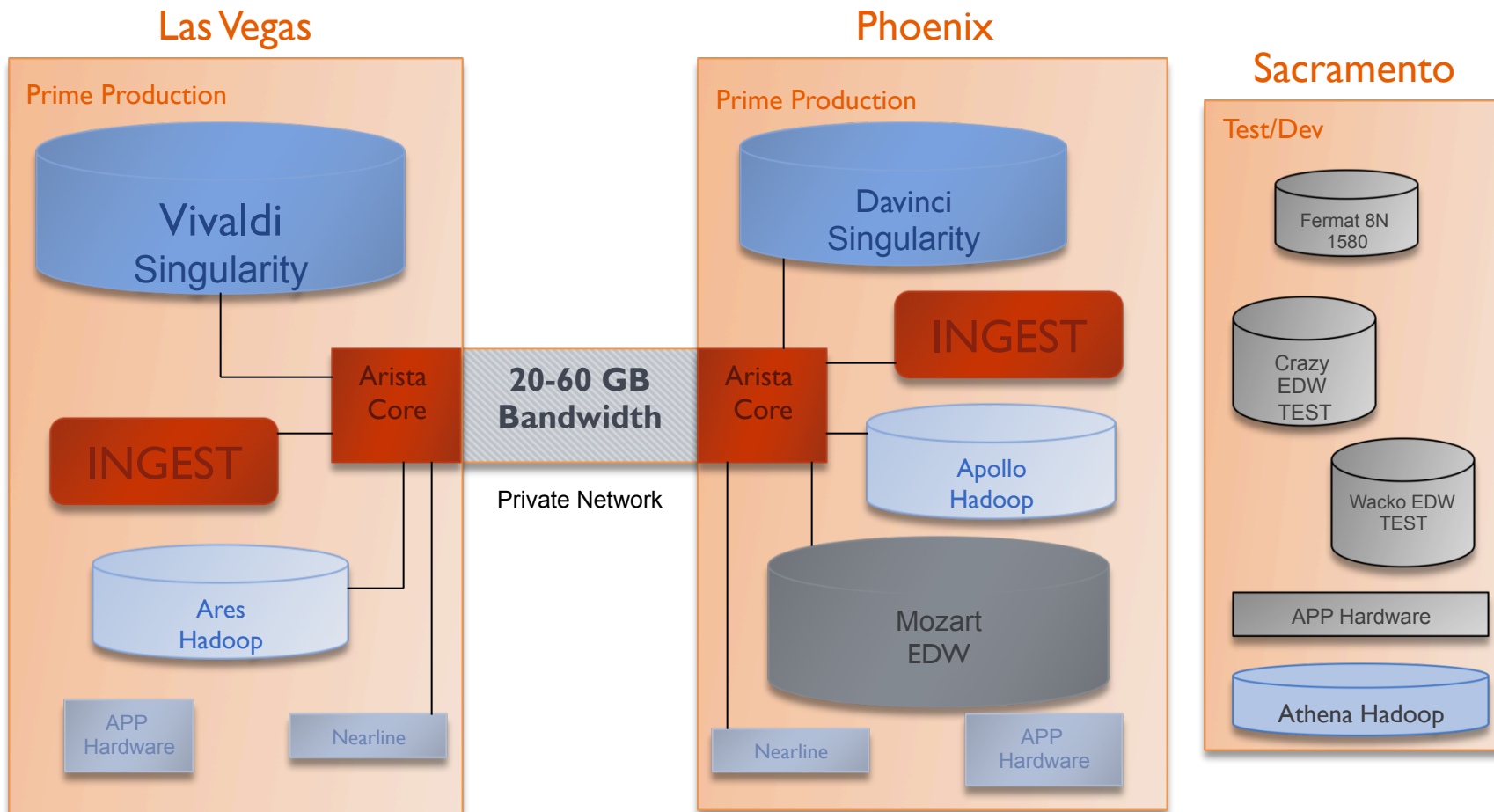
```
Teradata.Reader reader =  new
    Teradata.Reader(
        FileSystem.get(prodTD),
        mySQL
    );
While (reader.next(key, value)) {
…
}
```
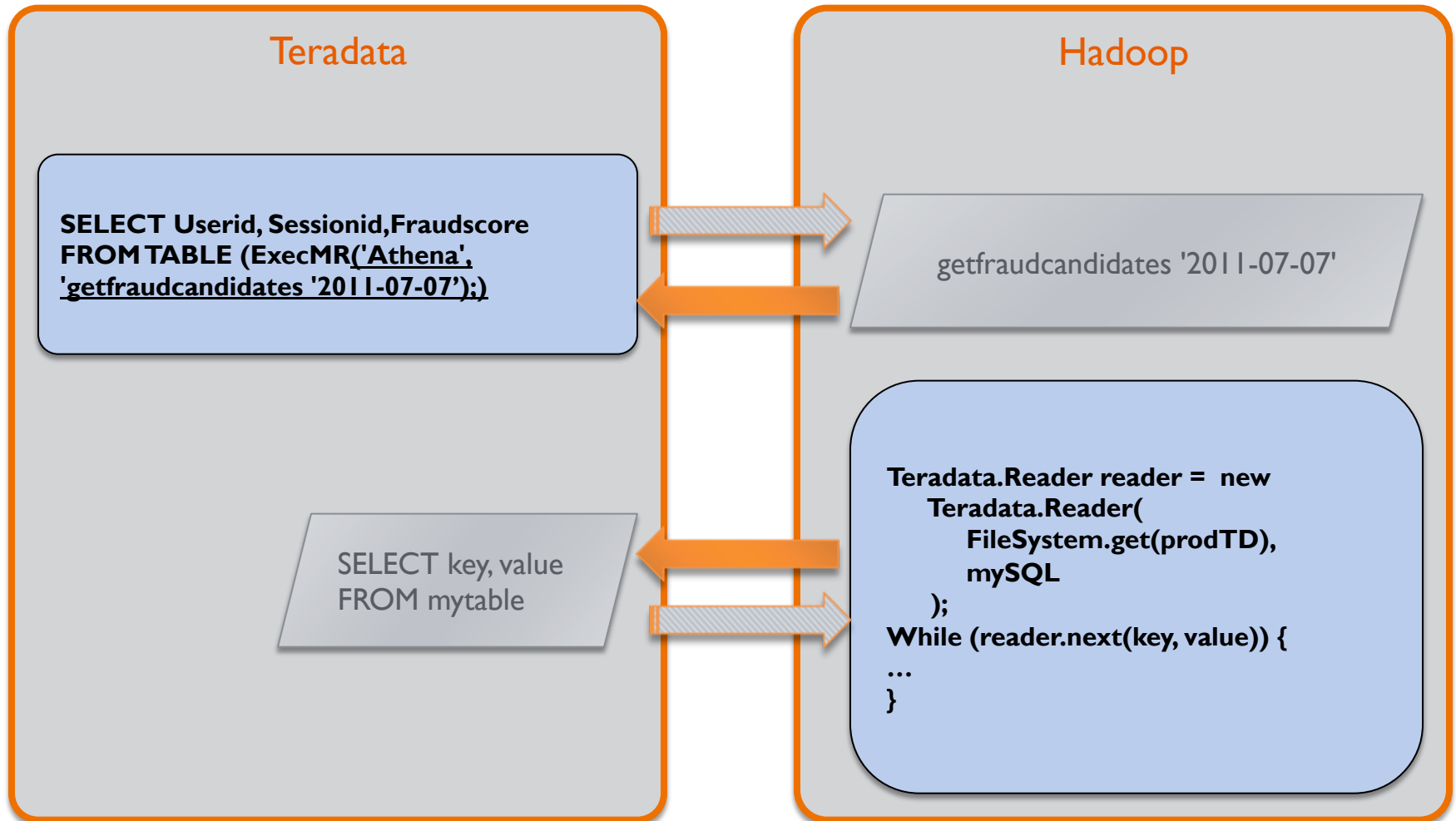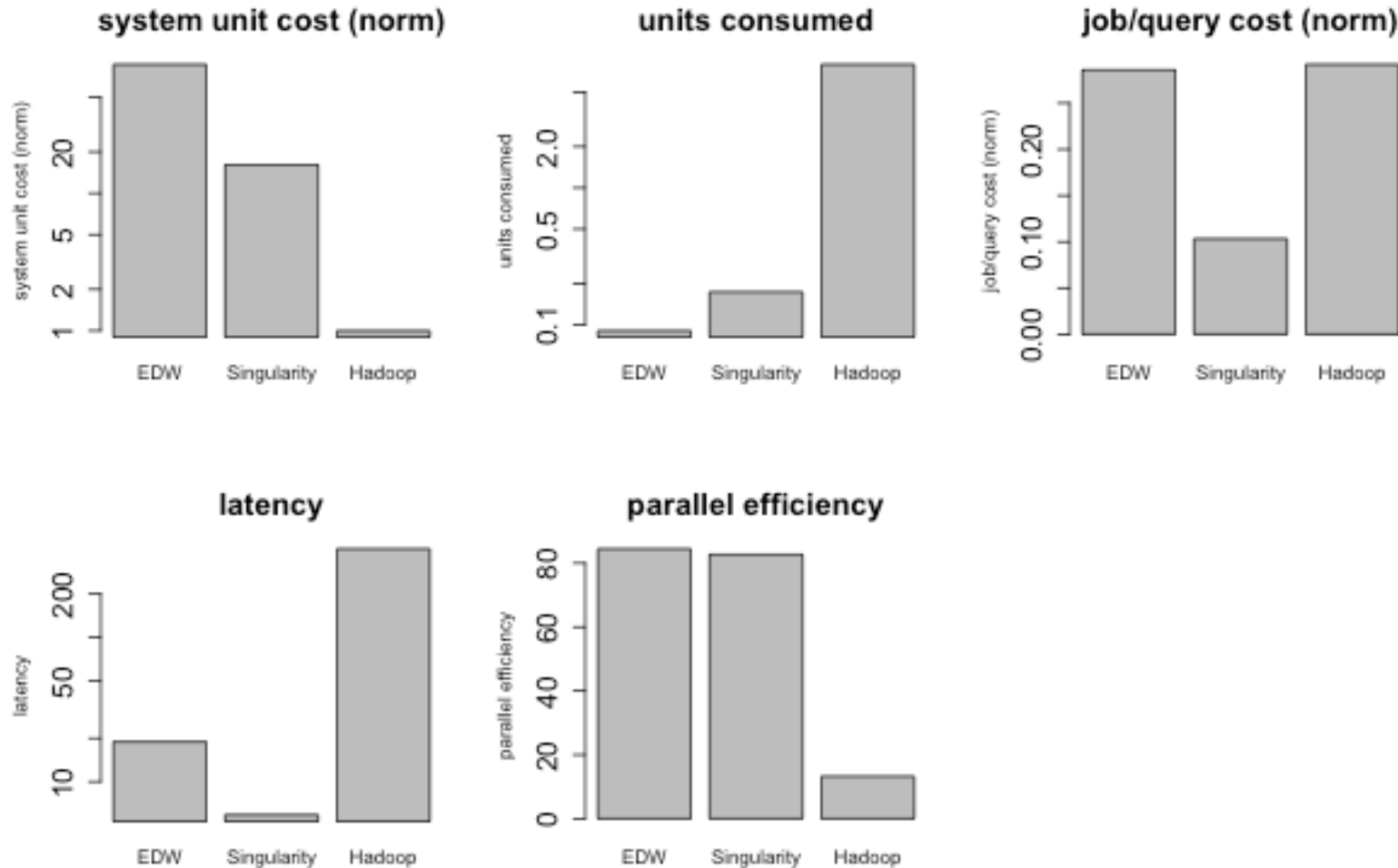
Platform Metrics for query 5

Table scan and summation

Analytics at eBay
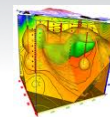
Predefined Early Binding Structured			Undefined Late Binding Unstructured
Columnar	Relational	Semi-Structured	Flat file
Cache/DataMart	Data Warehouse	Singularity	Hadoop