Extending the Enterprise Data Warehouse with Hadoop

Robert Lancaster

Nov 7, 2012

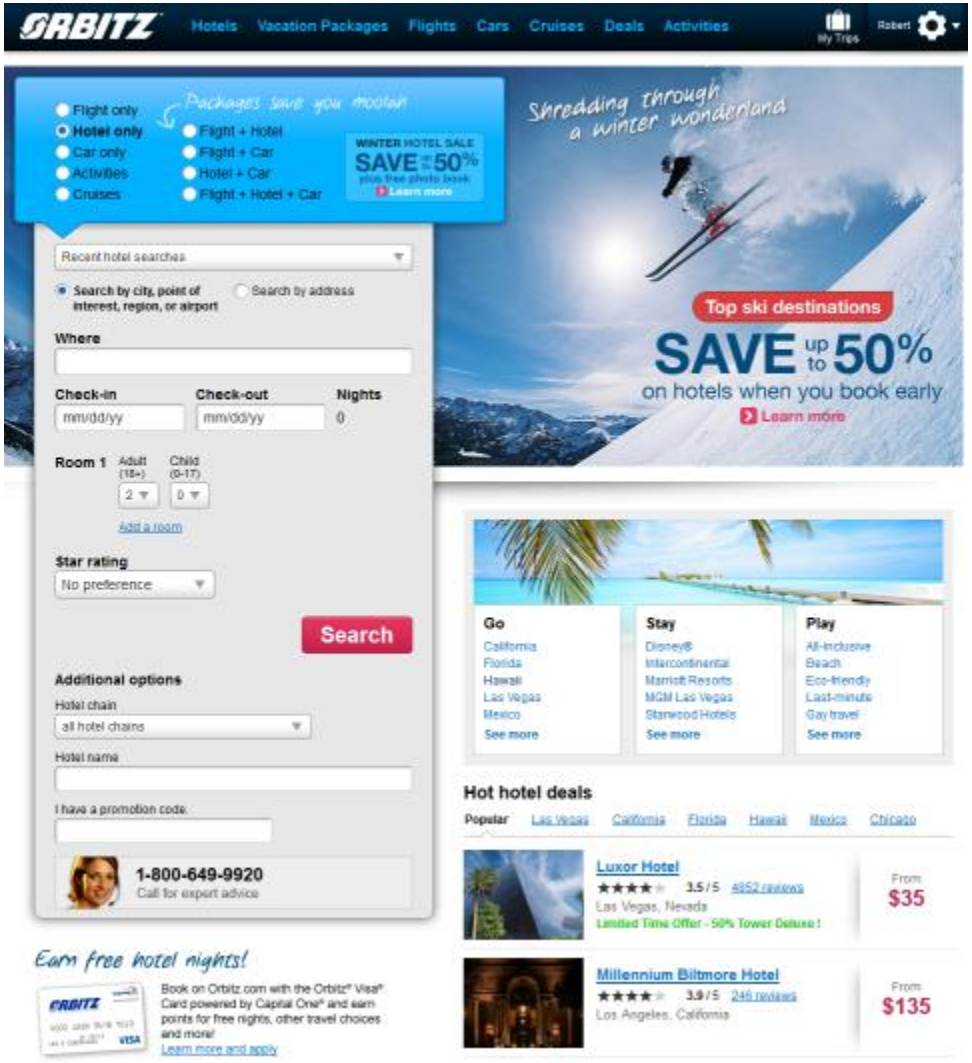
ORBITZ®

Who I Am

- Robert Lancaster
  - Solutions Architect, Hotel Supply Team
  - rlancaster@orbitz.com
  - @rob1lancaster
  - Organizer of Chicago Machine Learning Study Group
  - Co-organizer of Chicago Big Data.

Launched in 2001

Over 160 million bookings



○ Flight only
● Hotel only
○ Car only
○ Activities
○ Cruises

Packages save you moolah

○ Flight + Hotel
○ Flight + Car
○ Hotel + Car
○ Flight + Hotel + Car

WINTER HOTEL SALE
SAVE up to 50%
plus free photo book
Learn more

Recent hotel searches ▼

● Search by city, point of interest, region, or airport ○ Search by address

Where

Check-in mm/dd/yy Check-out mm/dd/yy Nights 0

Room 1 Adult (18+) 2 ▼ Child (0-17) 0 ▼

Add a room

Star rating
No preference ▼

Search

Additional options

Hotel chain
all hotel chains ▼

Hotel name

I have a promotion code:

1-800-649-9920
Call for expert advice

Shredding through a winter wonderland

Top ski destinations

SAVE up to 50%
on hotels when you book early
Learn more

Go	Stay	Play
California	Disney®	All-inclusive
Florida	Intercontinental	Beach
Hawaii	Marriott Resorts	Eco-friendly
Las Vegas	MGM Las Vegas	Last-minute
Mexico	Starwood Hotels	Gay travel
See more	See more	See more

Hot hotel deals

Popular Las Vegas California Florida Hawaii Mexico Chicago

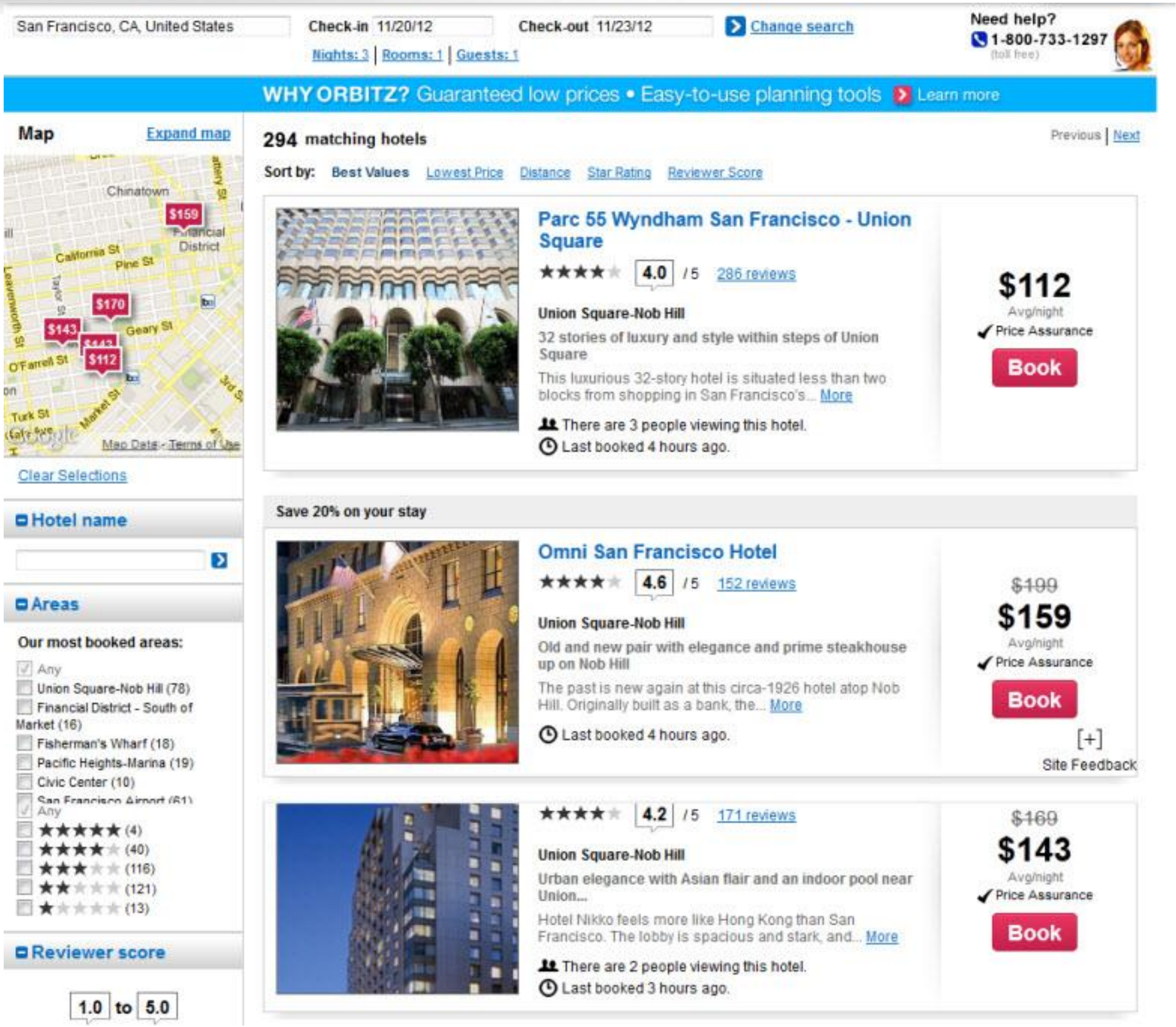
Luxor Hotel
★★★★☆ 3.5 / 5 4852 reviews
Las Vegas, Nevada
Limited Time Offer - 50% Tower Deluxe !
From $35

Millennium Biltmore Hotel
★★★★☆ 3.9 / 5 246 reviews
Los Angeles, California
From $135

Earn free hotel nights!

Book on Orbitz.com with the Orbitz® Visa® Card powered by Capital One® and earn points for free nights, other travel choices and more!
Learn more and apply

Some History…



San Francisco, CA, United States Check-in 11/20/12 Check-out 11/23/12 Change search

Nights: 3 | Rooms: 1 | Guests: 1

Need help?
☎ 1-800-733-1297
(toll free)

WHY ORBITZ? Guaranteed low prices • Easy-to-use planning tools Learn more

Map Expand map

Clear Selections

Hotel name

Areas

Our most booked areas:

☑ Any
□ Union Square-Nob Hill (78)
□ Financial District - South of Market (16)
□ Fisherman's Wharf (18)
□ Pacific Heights-Marina (19)
□ Civic Center (10)
□ San Francisco Airport (61)

☑ Any
□ ★★★★★ (4)
□ ★★★★☆ (40)
□ ★★★☆☆ (116)
□ ★★☆☆☆ (121)
□ ★☆☆☆☆ (13)

Reviewer score

1.0 to 5.0

294 matching hotels Previous | Next

Sort by: **Best Values** Lowest Price Distance Star Rating Reviewer Score

Parc 55 Wyndham San Francisco - Union Square
★★★★☆ 4.0 / 5 286 reviews
Union Square-Nob Hill
32 stories of luxury and style within steps of Union Square
This luxurious 32-story hotel is situated less than two blocks from shopping in San Francisco's... More
There are 3 people viewing this hotel.
Last booked 4 hours ago.
$112
Avg/night
✓ Price Assurance
Book

Save 20% on your stay

Omni San Francisco Hotel
★★★★☆ 4.6 / 5 152 reviews
Union Square-Nob Hill
Old and new pair with elegance and prime steakhouse up on Nob Hill
The past is new again at this circa-1926 hotel atop Nob Hill. Originally built as a bank, the... More
Last booked 4 hours ago.
~~$199~~
$159
Avg/night
✓ Price Assurance
Book
[+]
Site Feedback

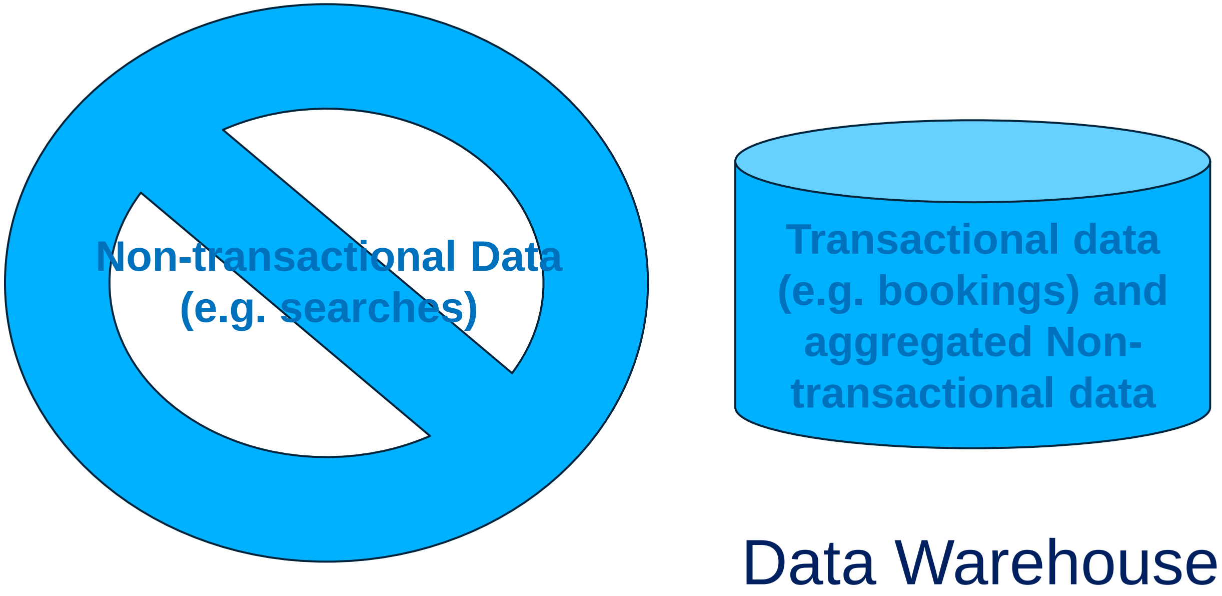
★★★★☆ 4.2 / 5 171 reviews
Union Square-Nob Hill
Urban elegance with Asian flair and an indoor pool near Union...
Hotel Nikko feels more like Hong Kong than San Francisco. The lobby is spacious and stark, and... More
There are 2 people viewing this hotel.
Last booked 3 hours ago.
~~$169~~
$143
Avg/night
✓ Price Assurance
Book

In 2009…

- The Machine Learning team is formed to improve site performance. For example, improving hotel search results.
- This required access to large volumes of behavioral data for analysis.
    - Fortunately, the required data was collected in session data stored in web analytics logs.

The Problem…

- The only archive of the required data went back about two weeks.

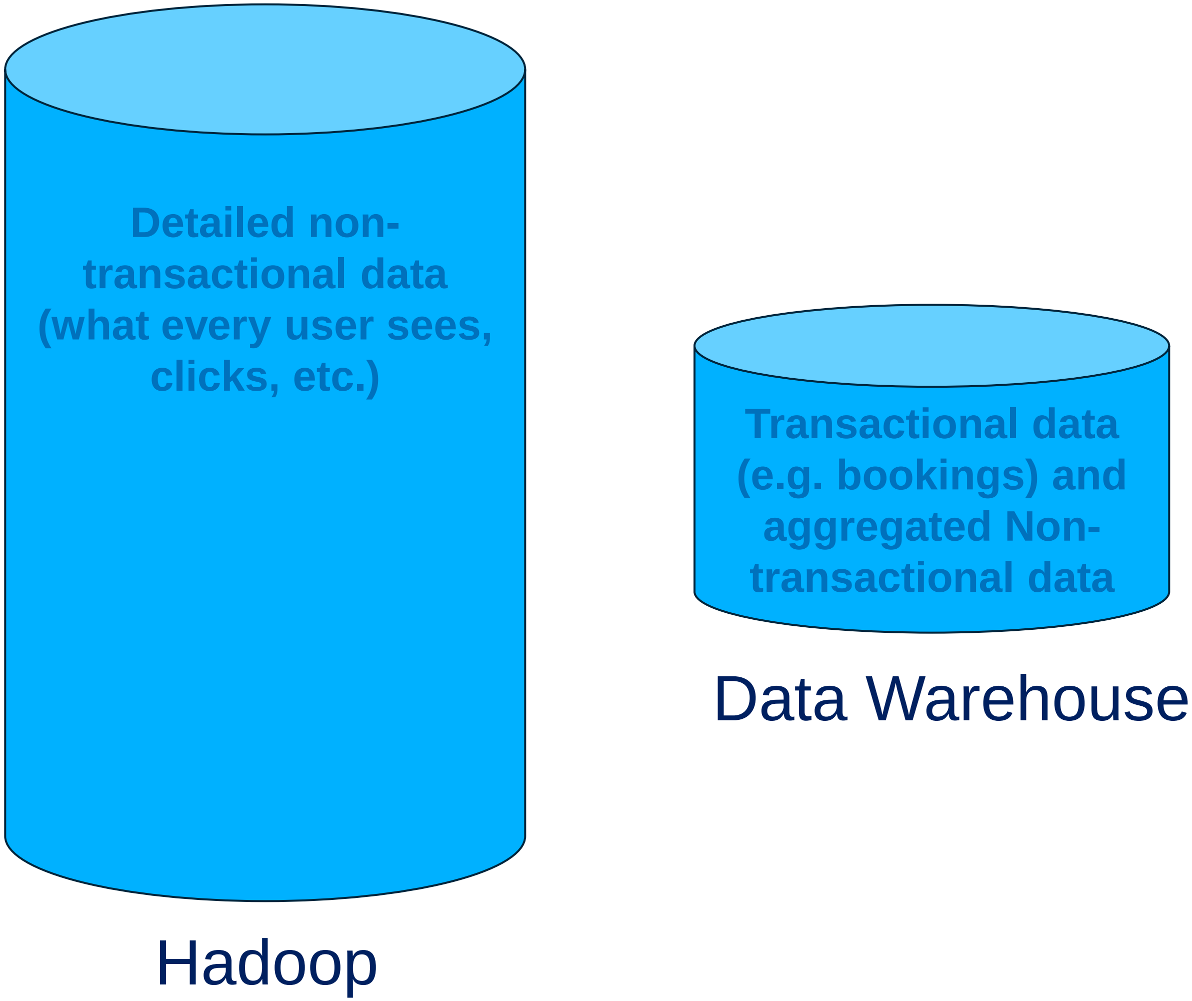


Non-transactional Data (e.g. searches)

Transactional data (e.g. bookings) and aggregated Non-transactional data

Data Warehouse

Hadoop Provided a Solution…



Detailed non-transactional data (what every user sees, clicks, etc.)

Hadoop

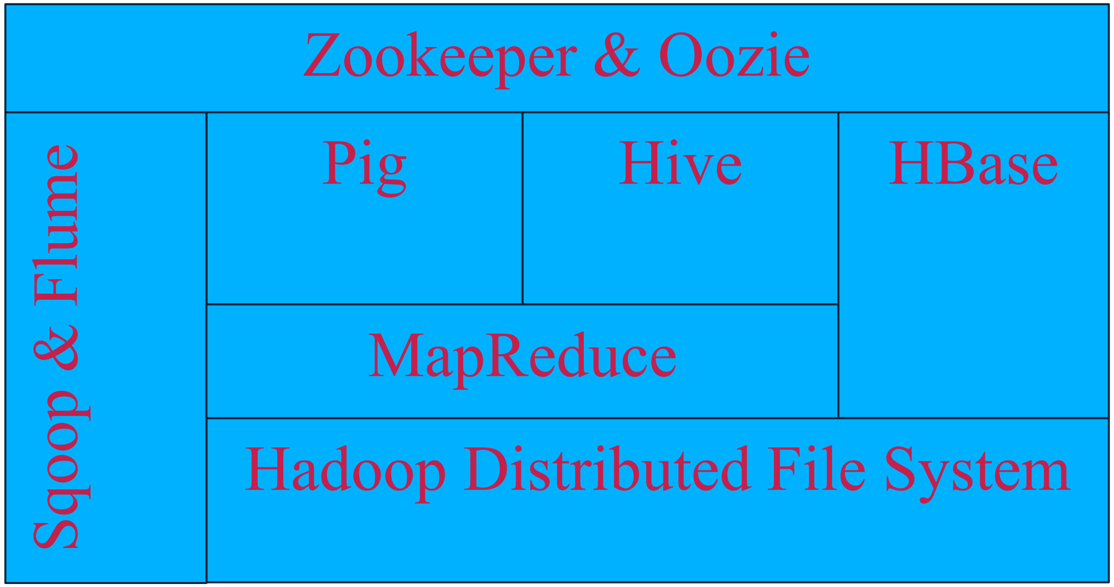
Transactional data (e.g. bookings) and aggregated Non-transactional data

Data Warehouse

What is Hadoop?

- Distributed file system and parallel processing platform.
- Open source Apache project created by Doug Cutting.
- Modeled on papers published by Google on the Google File System and MapReduce.
- Intended to run on a cluster of relatively inexpensive machines (aka commodity hardware).
- Bring processing to the data.

The Hadoop Ecosystem



Zookeeper & Oozie

Sqoop & Flume

Pig

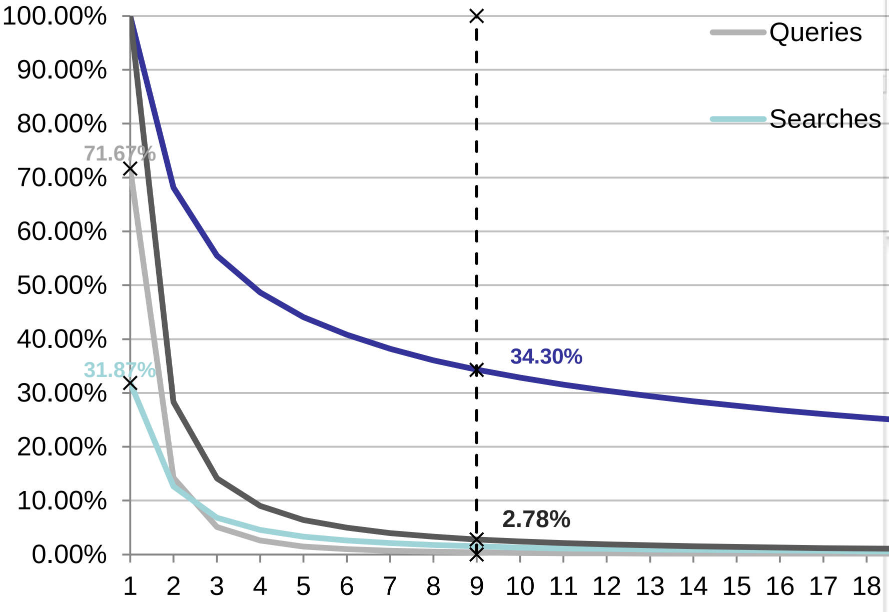
Hive

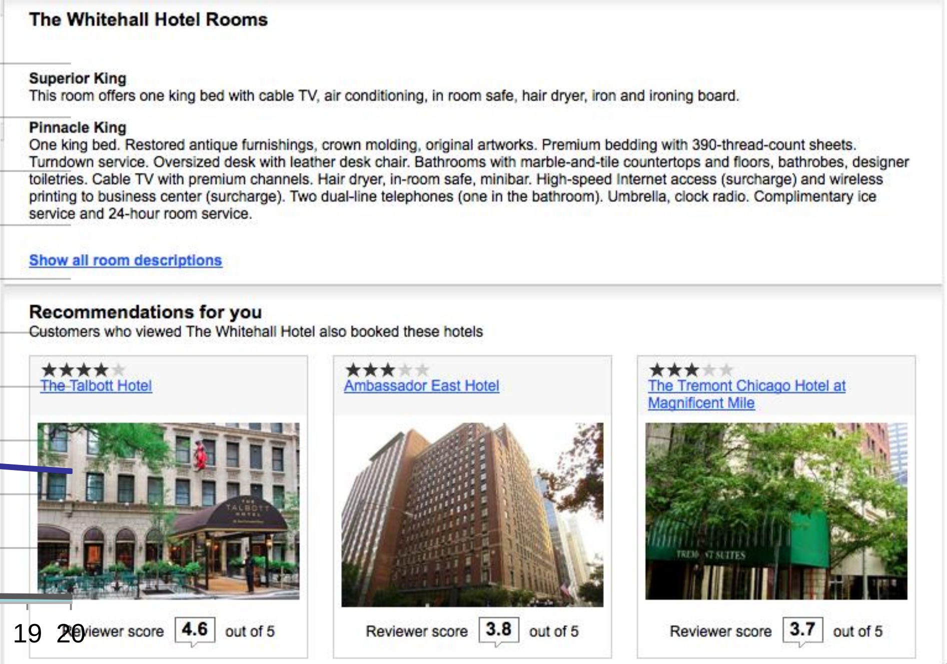
HBase

MapReduce

Hadoop Distributed File System

Deploying Hadoop Enabled Multiple Applications…





The Whitehall Hotel Rooms

Superior King
This room offers one king bed with cable TV, air conditioning, in room safe, hair dryer, iron and ironing board.

Pinnacle King
One king bed. Restored antique furnishings, crown molding, original artworks. Premium bedding with 390-thread-count sheets. Turndown service. Oversized desk with leather desk chair. Bathrooms with marble-and-tile countertops and floors, bathrobes, designer toiletries. Cable TV with premium channels. Hair dryer, in-room safe, minibar. High-speed Internet access (surcharge) and wireless printing to business center (surcharge). Two dual-line telephones (one in the bathroom). Umbrella, clock radio. Complimentary ice service and 24-hour room service.

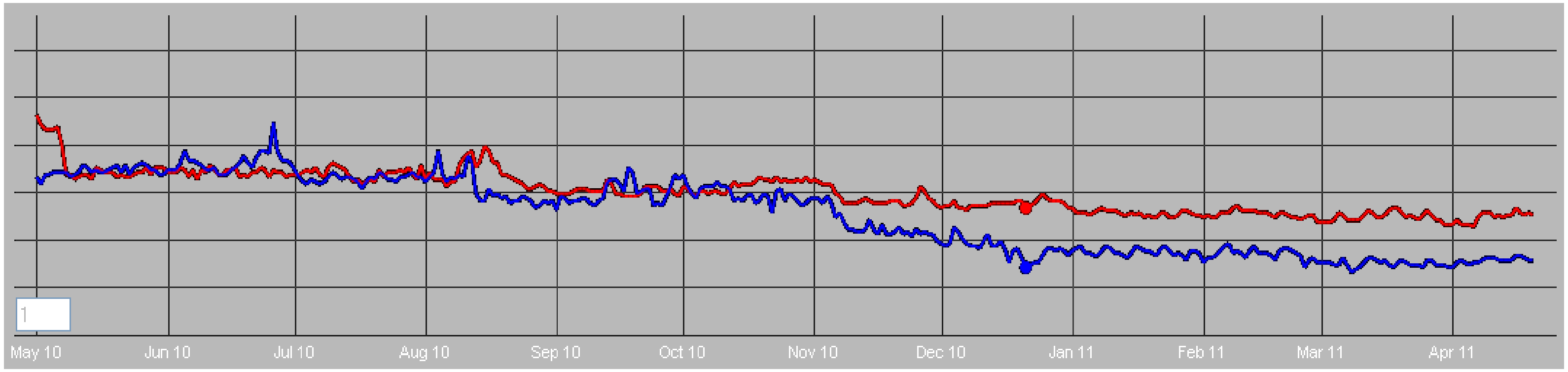
Show all room descriptions

Recommendations for you
Customers who viewed The Whitehall Hotel also booked these hotels

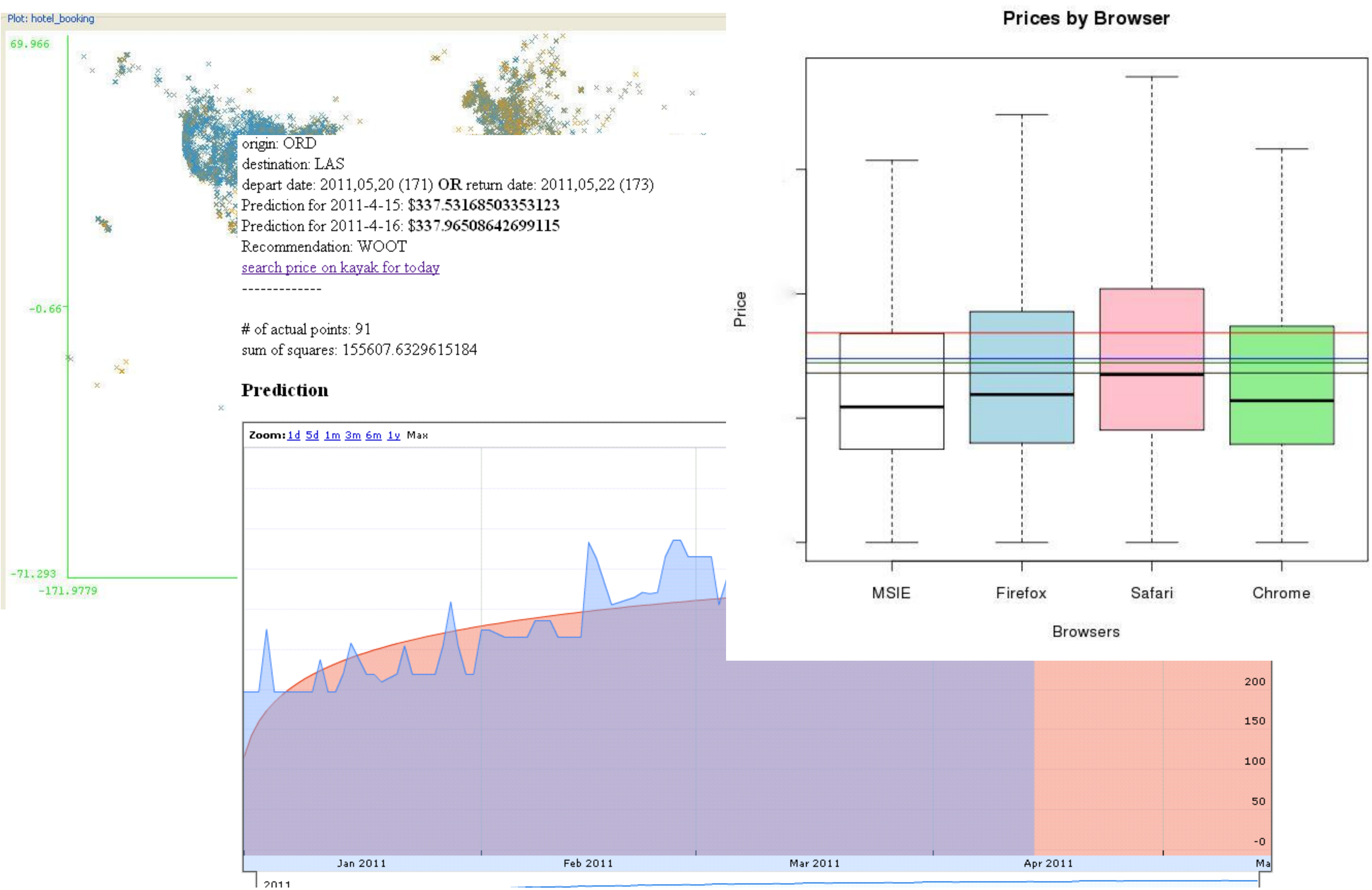
★★★★☆ The Talbott Hotel
Reviewer score 4.6 out of 5

★★★☆☆ Ambassador East Hotel
Reviewer score 3.8 out of 5

★★★☆☆ The Tremont Chicago Hotel at Magnificent Mile
Reviewer score 3.7 out of 5



And Useful Analyses…



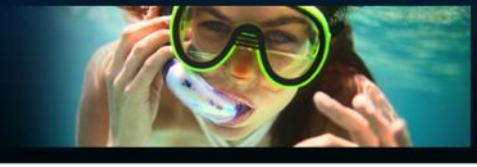
Plot: hotel_booking

origin: ORD
destination: LAS
depart date: 2011,05,20 (171) **OR** return date: 2011,05,22 (173)
Prediction for 2011-4-15: **$337.53168503353123**
Prediction for 2011-4-16: **$337.96508642699115**
Recommendation: WOOT
[search price on kayak for today]()

# of actual points: 91
sum of squares: 155607.6329615184

Prediction

But Brought New Challenges…



- Most of these efforts are driven by development teams.
- The challenge now is unlocking the value of this data for non-technical users.
- Support for Hadoop via traditional BI/reporting tools still meager.

BI Vendors Are Working on Hadoop Integration

Both big (relatively)…

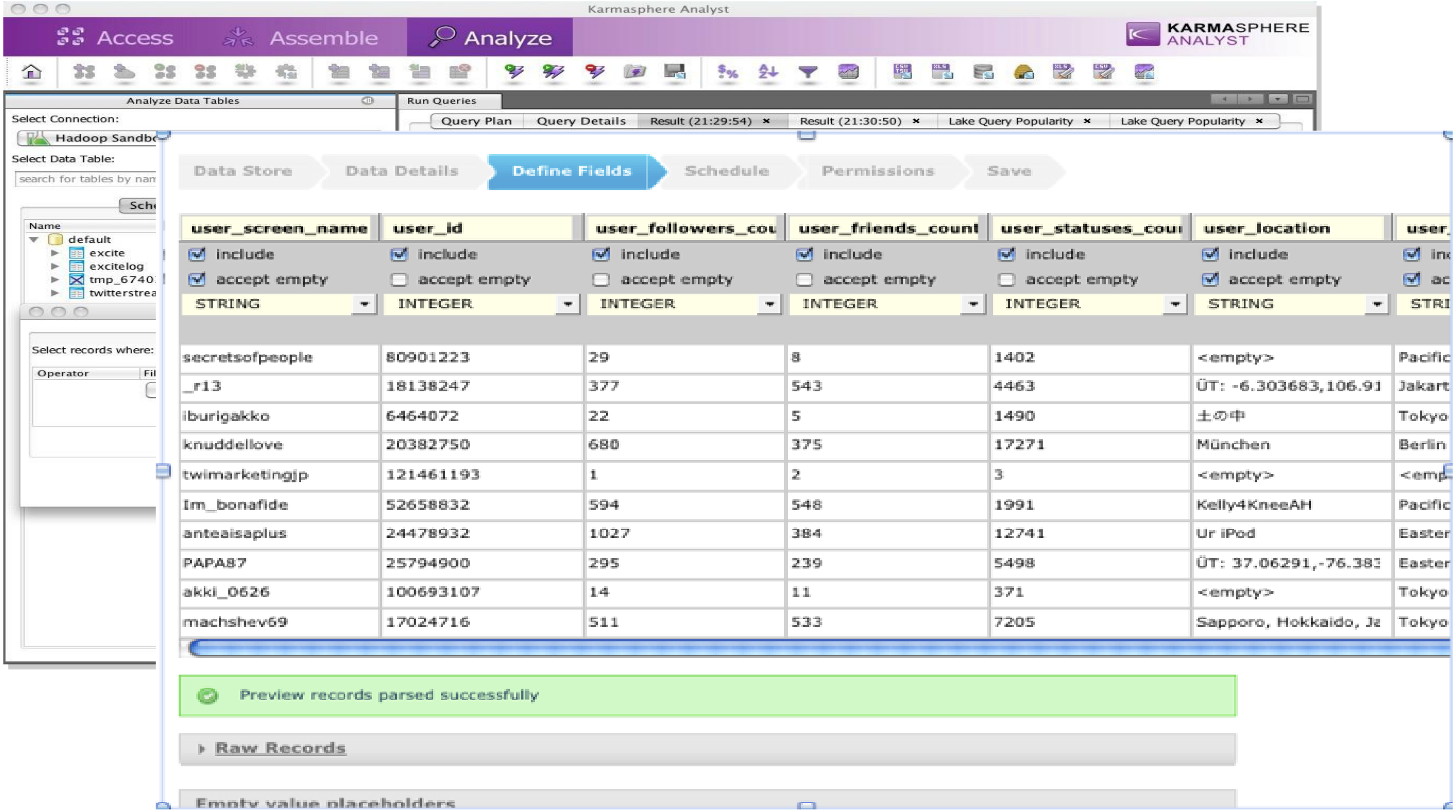
INFORMATICA
The Data Integration Company™

MicroStrategy

QlikView

GREENPLUM®

And small…



user_screen_name	user_id	user_followers_cou	user_friends_count	user_statuses_cou	user_location	user_
☑ include	☑ include	☑ include	☑ include	☑ include	☑ include	☑ inc
☑ accept empty	□ accept empty	□ accept empty	□ accept empty	□ accept empty	☑ accept empty	☑ ac
STRING	INTEGER	INTEGER	INTEGER	INTEGER	STRING	STRI
secretsofpeople	80901223	29	8	1402	<empty>	Pacific
_r13	18138247	377	543	4463	ÜT: -6.303683,106.91	Jakart
iburigakko	6464072	22	5	1490	土の中	Tokyo
knuddellove	20382750	680	375	17271	München	Berlin
twimarketingjp	121461193	1	2	3	<empty>	<emp
Im_bonafide	52658832	594	548	1991	Kelly4KneeAH	Pacific
anteaisaplus	24478932	1027	384	12741	Ur iPod	Easter
PAPA87	25794900	295	239	5498	ÜT: 37.06291,-76.383	Easter
akki_0626	100693107	14	11	371	<empty>	Tokyo
machshev69	17024716	511	533	7205	Sapporo, Hokkaido, Ja	Tokyo

Preview records parsed successfully

Raw Records

Empty value placeholders

In 2011& 2012

- Big Data team is formed under Business Intelligence team at Orbitz Worldwide.
- Allows the Big Data team to work more closely with the data warehouse and BI teams.
- Reflects the importance of big data to the future of the company.
- Our production cluster has grown 40-fold since it was launched.

A View Shared Beyond Orbitz…

“We strongly believe that Hadoop is the nucleus of the next-generation cloud EDW…”

“…but that promise is still three to five years from fruition.”*

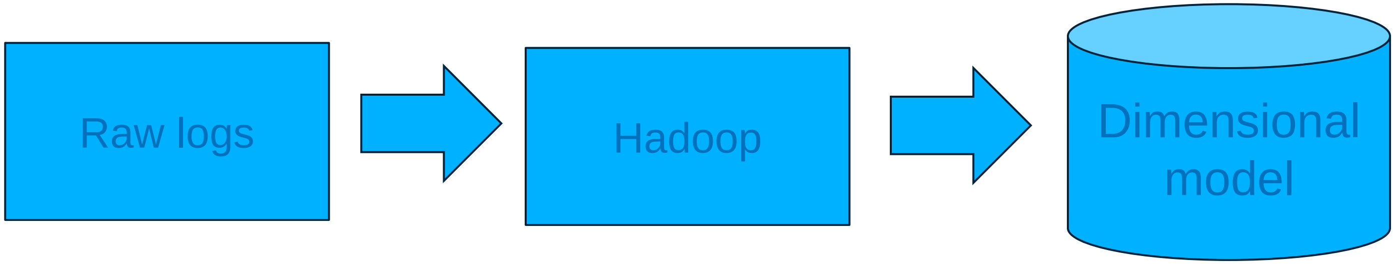
*James Kobielus, Forrester Research,
“Hadoop, Is It Soup Yet?”

Two Primary Ways We Use Hadoop to Complement the EDW

- Extraction and transformation of data for loading into the data warehouse – “ETL”.
- Off-loading of analysis from the data warehouse.

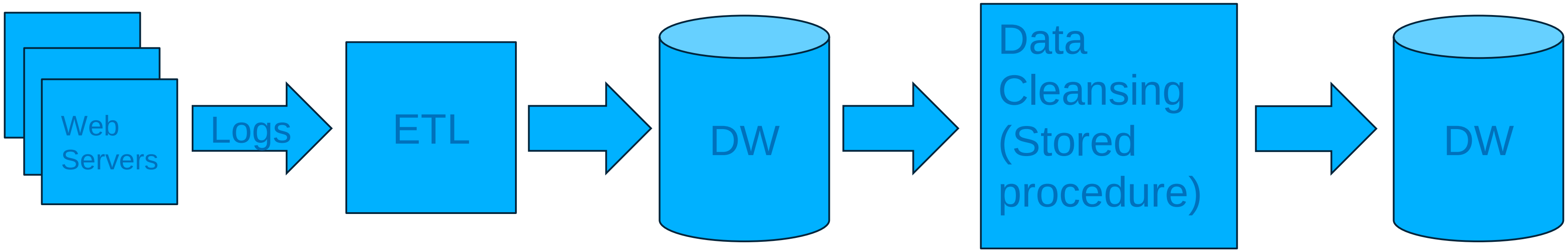
ETL Example

Proposed Processing



ETL Example: Click Data Processing

Previous Processing in Data Warehouse

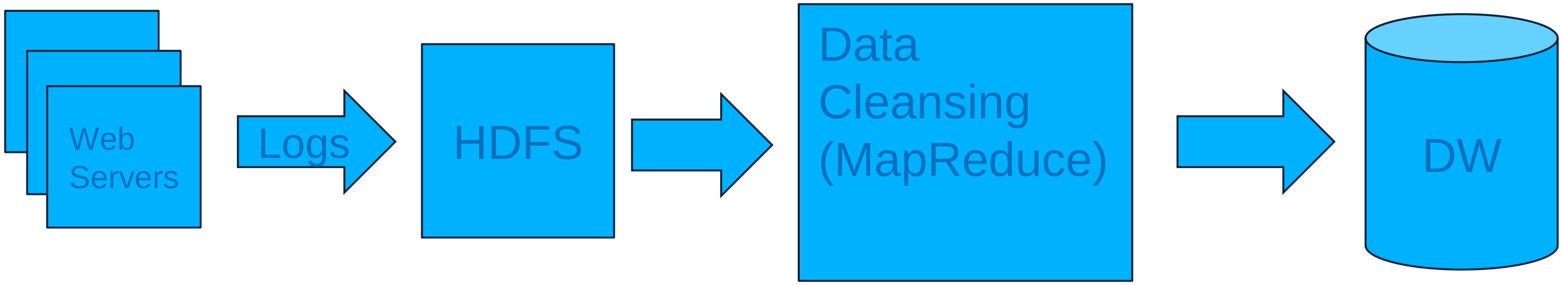


Several hours of processing

~20% original data size

ETL Example: Click Data Processing

- Moving to Hadoop:
  - Removed load from the data warehouse.
  - Facilitated adding additional attributes for processing.
  - Allowed processing to be run more frequently.



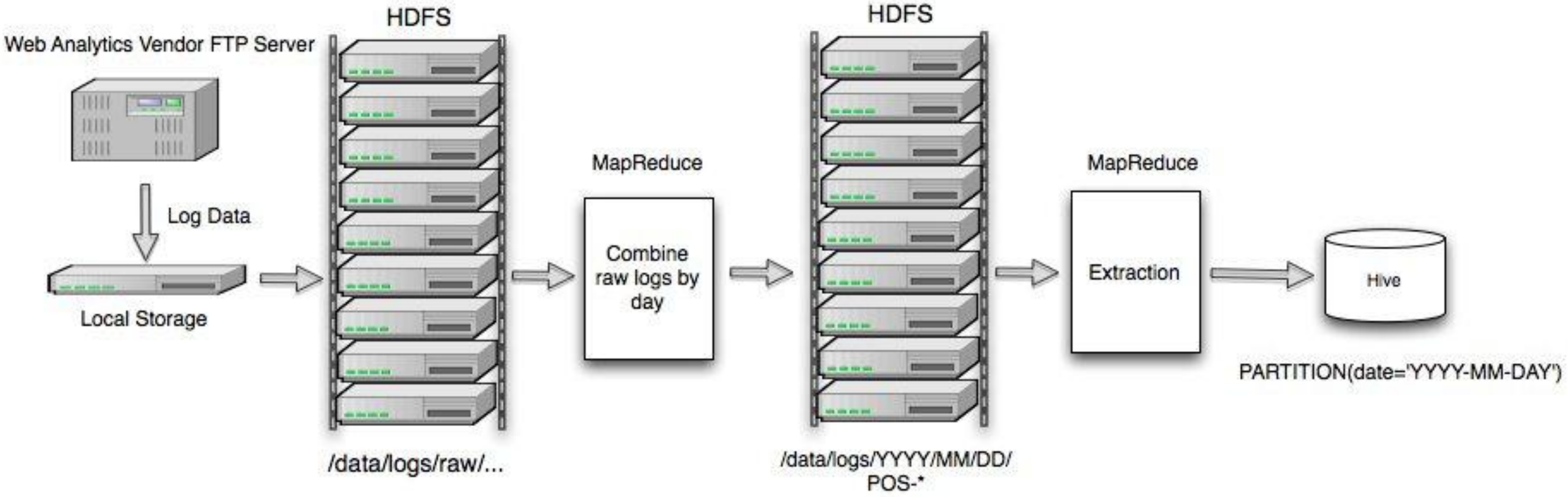
Processing in Hadoop

Analysis Example: Geo-Targeting Ads

- Facilitated analysis that allows for more personalized ad content.
- Allowed marketing team to analyze over a years worth of search data.
- Provided analysis that was difficult to perform in the data warehouse.

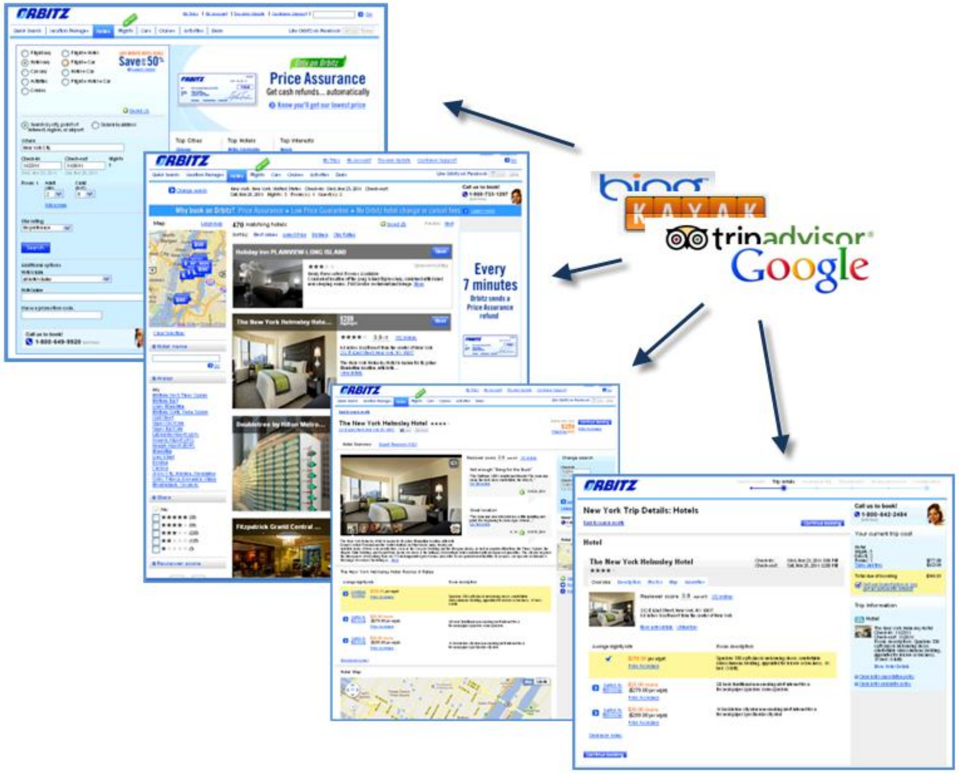
Looking for
CHEAP
AIRFARES?
Find Deals
Compare more sites with just one click.
trip.com

Example Processing Pipeline for Web Analytics Data



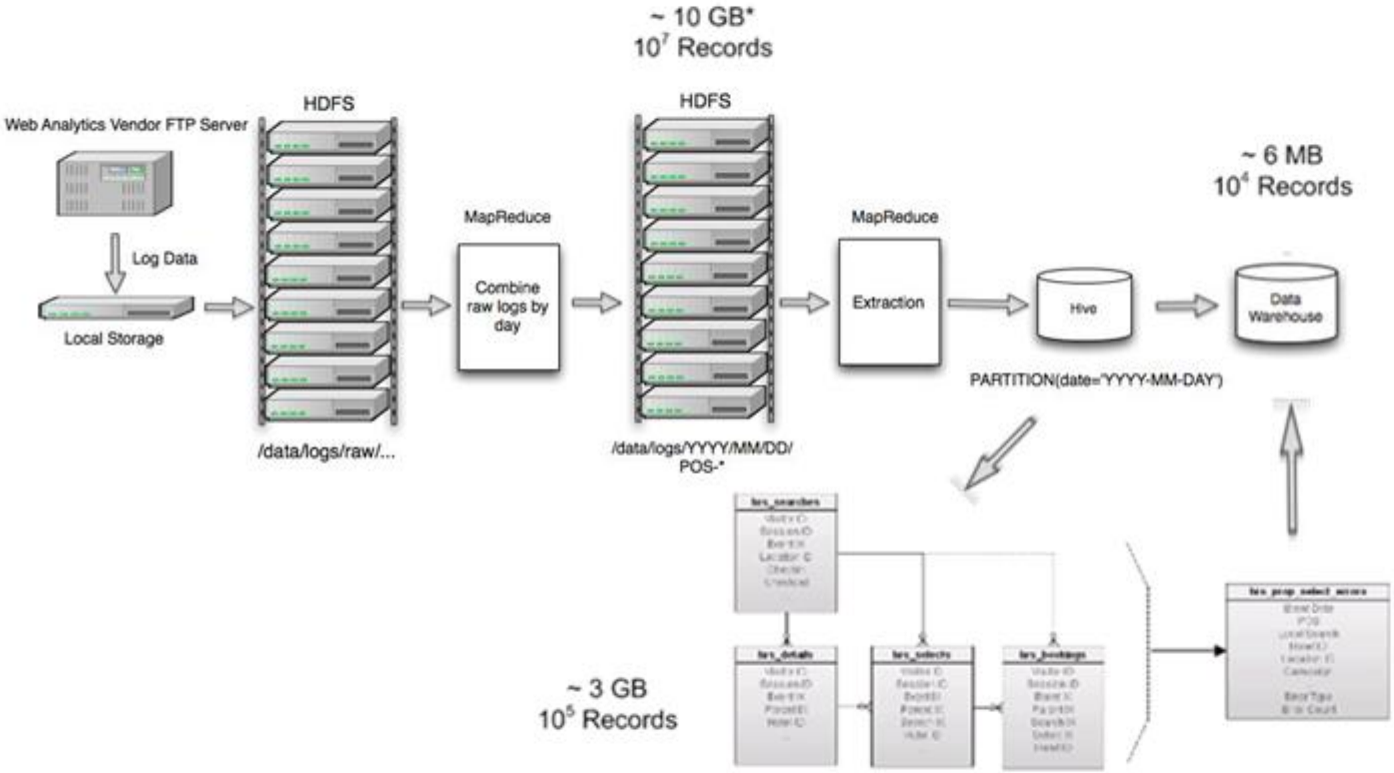
Example Use Case: Selection Errors

Use Case – Selection Errors: Introduction

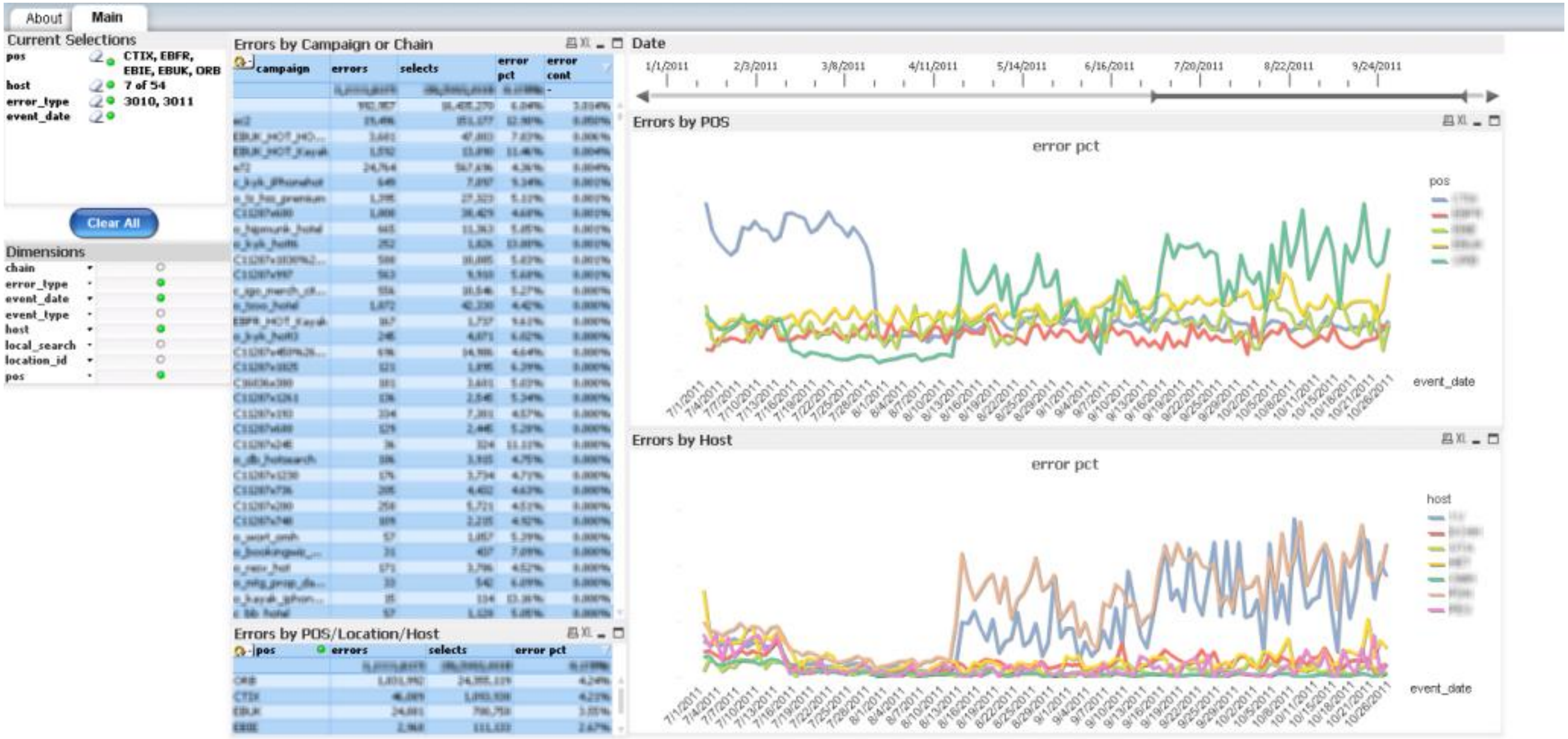

ORBITZ
Save 50%
Price Assurance
Get cash refunds... automatically
Every 7 minutes
Orbitz sends a Price Assurance refund
bing
KAYAK
tripadvisor
Google
New York Trip Details: Hotels
The New York Helmsley Hotel

- Multiple points of entry.
- Multiple paths through site.
- Goal: tie events together to form picture of customer behavior.

Use Case – Selection Errors: Processing



Use Case – Selection Errors: Visualization



Current Selections

pos	CTIX, EBFR, EBIE, EBUK, ORB
host	7 of 54
error_type	3010, 3011
event_date	

Dimensions

chain	○
error_type	●
event_date	●
event_type	○
host	●
local_search	○
location_id	○
pos	●

Errors by Campaign or Chain

campaign	errors	selects	error pct	error cont
	[redacted]	[redacted]	[redacted]	-
	992,957	16,435,270	6.04%	3.034%
ec2	19,496	151,177	12.90%	0.050%
EBUK_HOT_HO...	3,601	47,003	7.67%	0.006%
EBUK_HOT_Kayak	1,592	13,890	11.46%	0.004%
ef2	24,764	567,636	4.36%	0.004%
c_kyk_iPhonehot	649	7,097	9.14%	0.002%
o_lb_hot_premium	1,395	27,323	5.11%	0.002%
C11287v600	1,808	38,429	4.68%	0.001%
o_hipmunk_hotel	645	11,363	5.68%	0.001%
o_kyk_hotl6	252	1,826	13.80%	0.001%
C11287v3030%2...	588	10,085	5.83%	0.001%
C11287v997	563	9,910	5.68%	0.001%
c_gpo_merch_ot...	556	10,546	5.27%	0.000%
o_trivo_hotel	1,872	42,330	4.42%	0.000%
EBFR_HOT_Kayak	167	1,757	9.41%	0.000%
o_kyk_hotl3	246	4,071	6.02%	0.000%
C11287v48O%26...	696	14,986	4.64%	0.000%
C11287v1025	121	1,895	6.39%	0.000%
C36036a300	181	3,601	5.03%	0.000%
C11287v1261	136	2,546	5.34%	0.000%
C11287v193	334	7,301	4.57%	0.000%
C11287v600	129	2,446	5.29%	0.000%
C11287v248	36	324	11.11%	0.000%
o_db_hotsearch	186	3,915	4.75%	0.000%
C11287v1230	176	3,734	4.71%	0.000%
C11287v736	205	4,402	4.63%	0.000%
C11287v200	258	5,721	4.51%	0.000%
C11287v748	109	2,215	4.92%	0.000%
o_wort_omh	57	1,057	5.39%	0.000%
o_bookingwiz_...	31	437	7.09%	0.000%
o_nav_hot	171	3,796	4.52%	0.000%
o_mkg_prop_da...	33	542	6.09%	0.000%
o_kayak_iphon...	15	114	13.16%	0.000%
c_bb_hotel	57	1,120	5.09%	0.000%

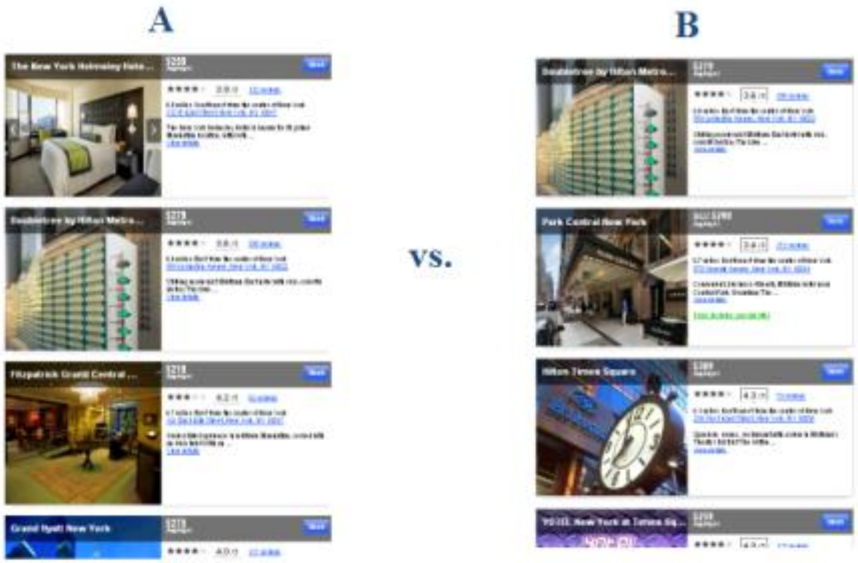
Errors by POS/Location/Host

pos	errors	selects	error pct
	[redacted]	[redacted]	[redacted]
ORB	1,031,992	24,355,119	4.24%
CTIX	46,089	1,093,938	4.21%
EBUK	24,881	700,758	3.55%
EBIE	2,968	111,133	2.67%

ORBITZ®

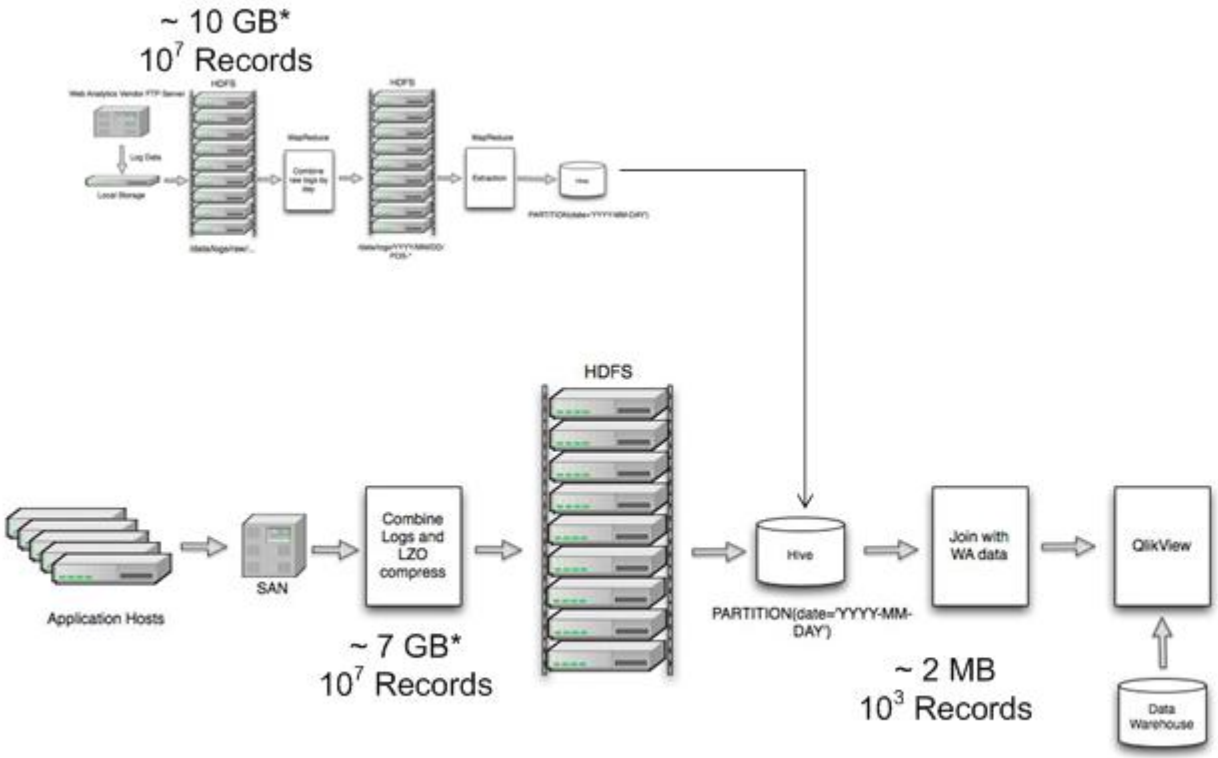
Example Use Case: Beta Data

Use Case – Beta Data: Introduction


A
B
vs.
The New York Helmsley Hote...
Doubletree by Hilton Metro...
Fitzpatrick Grand Central ...
Grand Hyatt New York
Doubletree by Hilton Metro...
Park Central New York
Hilton Times Square
YOTEL New York at Times Sq...

- Hotel Sort Optimization
- Compare A vs. B
- Web Analytics Data
  - What user saw.
  - How user behaved
- Server Log Data
  - Sorting behavior used.

Use Case – Beta Data Processing

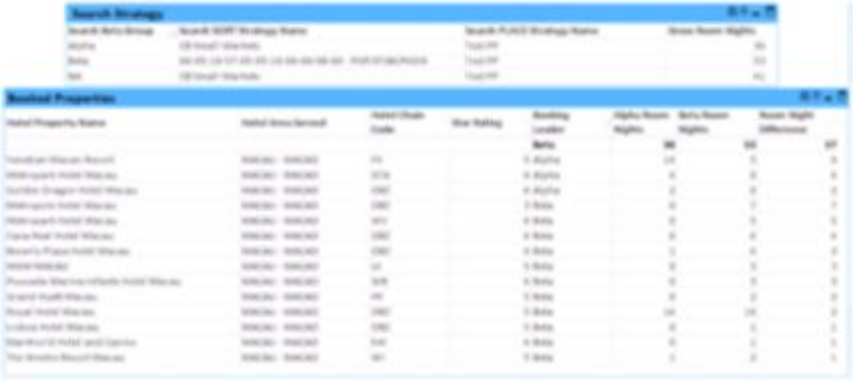


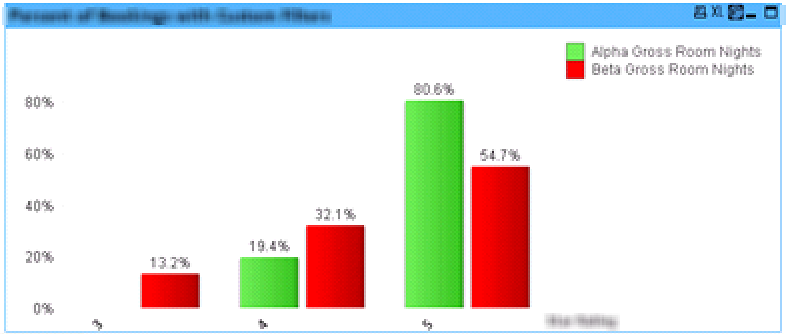
~ 10 GB*
10^7 Records

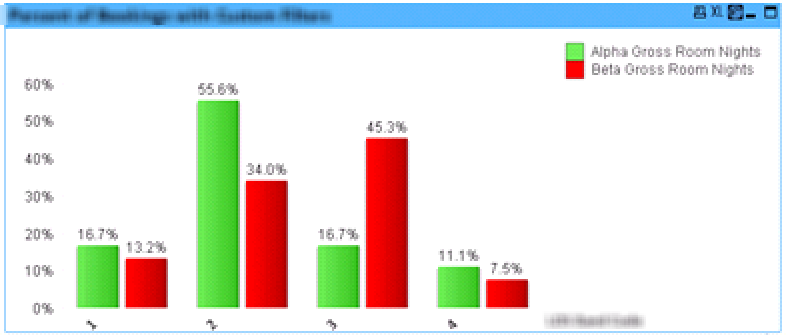
~ 7 GB*
10^7 Records

~ 2 MB
10^3 Records

Use Case – Beta Data: Visualization





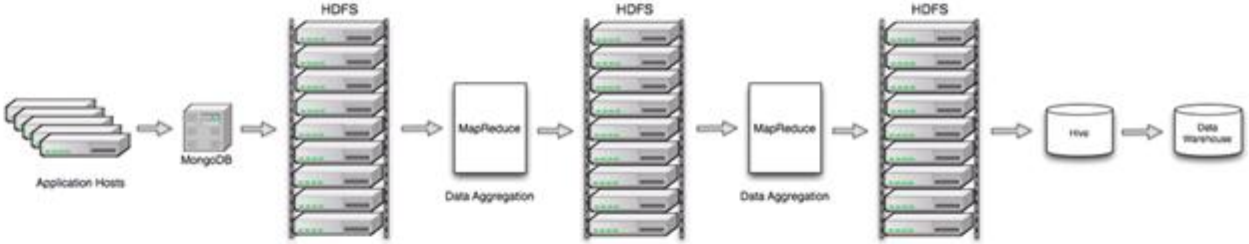


Example Use Case: RCDC

Use Case – RCDC: Introduction

- Understand and improve cache behavior.
- Improve “coverage”
  - Traditionally search 1 page of hotels at a time.
  - Get “just enough” information to present to consumers.
  - Increase amount of availability information we have when consumer performs a search.
- Data needed to support needs beyond reporting.

Use Case – RCDC: Processing

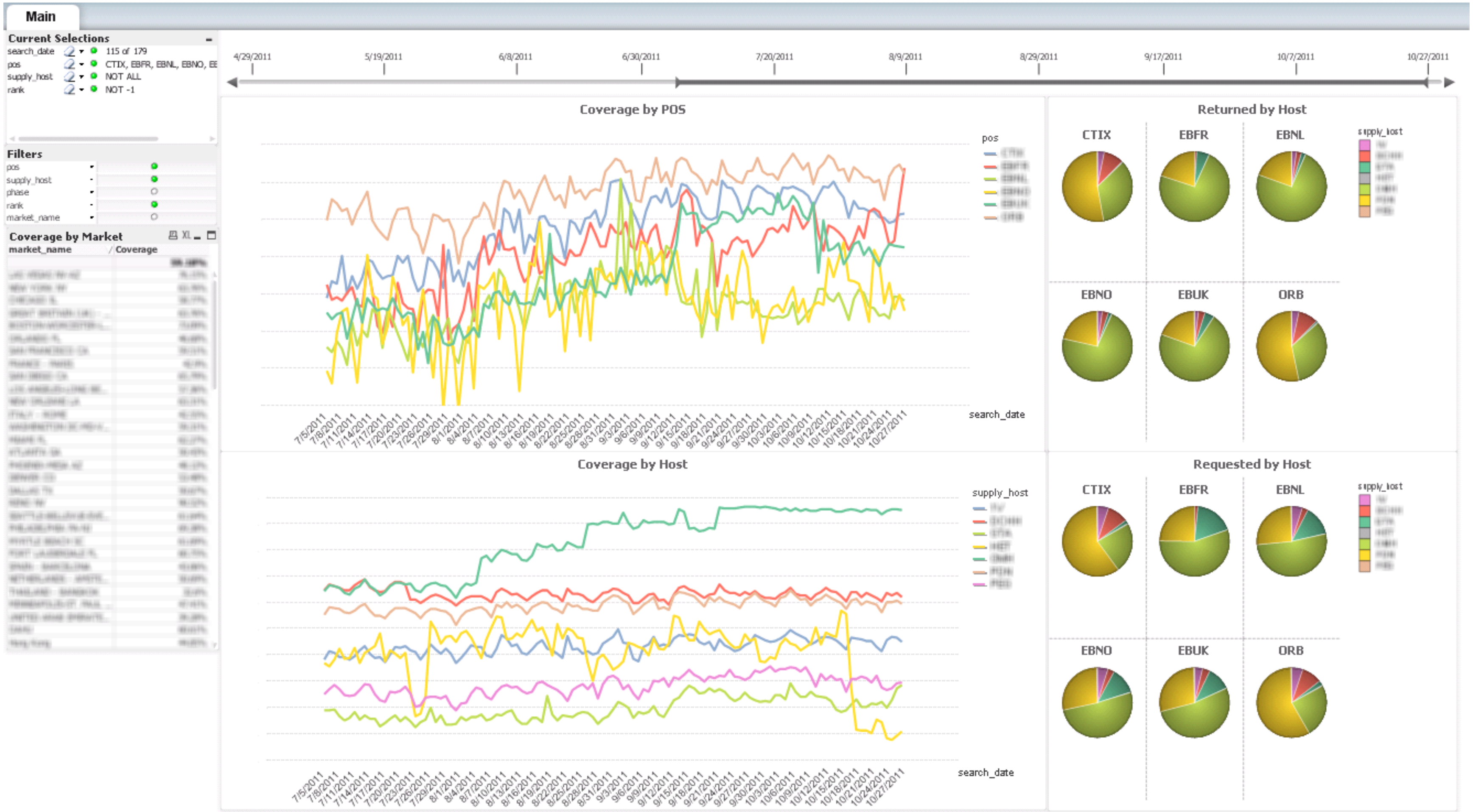


~ 500 GB
10^8 Records

~ 5 GB
10^5 Records

~ 5 MB
10^4 Records

Use Case – RCDC: Visualization



Main

Current Selections

search_date	115 of 179
pos	CTIX, EBFR, EBNL, EBNO, EE
supply_host	NOT ALL
rank	NOT -1

Filters

- pos ●
- supply_host ●
- phase ○
- rank ●
- market_name ○

Coverage by Market

market_name	Coverage
	[illegible]

Conclusions

- Hadoop market is still immature, but growing quickly. Better tools are on the way.
  - Look beyond the usual (enterprise) suspects. Many of the most interesting companies in the big data space are small startups.
- Hadoop won’t replace your EDW, but any organization with a large EDW should at least be exploring Hadoop as a complement to their BI infrastructure.

Conclusions

- Work closely with your existing data management teams.
  - Your idea of what constitutes “big data” might quickly diverge from theirs.
- The flip-side to this is that Hadoop can be an excellent tool to off-load resource-consuming jobs from your data warehouse.

ORBITZ

Thank you!

Questions?