



BI on Cloud Computing

Rohit Chatter

BI & Data Architect

Data & Insights Group

Yahoo India R & D, Bangalore



Agenda

- Cloud Computing – Quick look
- Cloud@Yahoo – The Yahoo! way
- Case study of BI on Cloud @ Yahoo – Live case
- My personal views – sharing experience
- Q & A



Cloud Computing

- What is it?
 - style of computing where massively scalable IT-related capabilities are provided “as a service” using Internet technologies to multiple external customers
 - E.g. Amazon EC2/S3, Yahoo, Google
- Key Features
 - Multi-Tenancy, On demand resources, Device & location independence, API based, scalability and others
- A Perspective

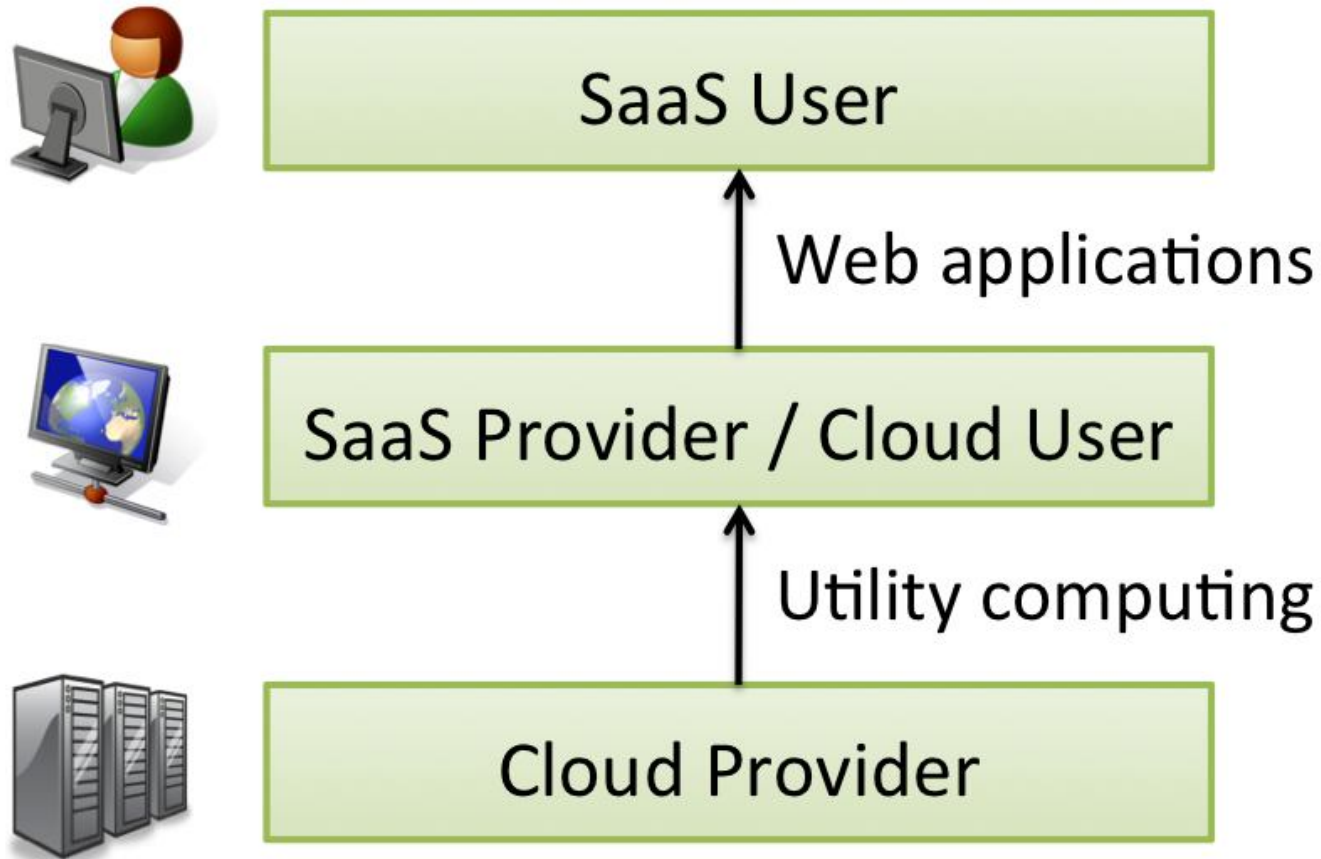


Focus on the business!

- Desire - start a new website, **iwanttosell.com**
- Service - provide listings of items for sale, jobs, etc.
- Business does well and more features needed
 - How do we scale for demand?
- Store listings as (key, category, description)
- Customers quickly ask for keyword search
- Add photos to listings
- And the business continues to grow and grow!



Cloud Computing – a perspective



• Copyright University of California at Berkeley



Internet Scale Generates *BigData*

- Yahoo is the most Visited Site on the Internet
 - 600M+ Unique Visitors per Month
 - Billions of Page Views per Day
 - Billions of Searches per Month
 - Billions of Emails per Month
 - **Terabytes of Data per Day!**
- And we crawl the Web
 - 100+ Billion Pages
 - 5+ Trillion Links
 - **Petabytes of data**
- Reading 100 Terabytes could be overwhelming

Std PC – 100Mbps	Server – 10Gbps	1000 Std PC
~ 11 days	~ 1 day	~ 15 mins



How is Yahoo seeing the space?

- **Yahoo sees two kinds of cloud services:**
 - Horizontal Cloud Services
 - Functionality enabling tenants to build applications or new services on top of the cloud
 - The focus of CCDI
 - Functional Cloud Services
 - Functionality that is useful in and of itself to tenants.
 - Yahoo!'s IndexTools; Yahoo! properties aimed at end-users e.g., flickr, Groups, Mail, News, Shopping
 - Could be build on top of horizontal cloud services or from scratch
- **Technology – Open Source adoption**
 - Hadoop – Grid
 - PIG – Programming language
 - ZooKeeper -- High-Availability Directory and Configuration Service
 - Oozie – Workflow engine



BI on Cloud – Case study

- Motivation
 - Report & Data requirements unknown
 - Evolving needs
 - Large data processing on demand
 - Web based access
- Architecture
- Functional View
- What is computed where?
- Few screenshots
- Benefits



BI on Cloud – Architecture



Functional View

Apache Web Server
PHP

Microstrategy/Home
e Grown

Oracle RDBMS
BI Aggregates
(H,D,W,M)

Utility Computing
Build Aggregates

Data Source
Dimension & Fact

Load balanced web

App Server – BI layer

Aggregates &
Metadata layer

Hadoop Grid + PIG
Cloud

Data – 100+ Gigabytes/Day

What is computed where

Derived Metrics – CTR, Depth,
RPM, Coverage

Rollups, Type 2 Dimension,
Alerts & Messaging

Metrics

Impressions, Revenue, Clicks,
Conversions, Quality Score,
Top keywords



BI on Cloud – Screenshot

[Logout](#)

test_svs

Report Preview

Ad Unit Id	Ad Unit Name	Account Id	Bidded Clicks	Bidded Searches	Broad Revenue Pub Cur
No records found.					

Filter List

Name	Operator	Input Type	Value	And/OR
startDate	equal to	date	1/1/2011	and
endDate	equal to	date	1/18/2011	and
Bidded Clicks	greater than	text	0	or
Bidded Searches	equal to	text		or
Broad Revenue Pub Cur	equal to	text		or

Requesting User: sshishir

Pig Script

```

R1_dim_msft_ad_unit_snapshot = LOAD '/projects/apollo/dim_msft_ad_unit_snapshot/daily/data/20110118' USING PigStorage('\u0001');
R2_dim_msft_ad_unit_snapshot = FOREACH R1_dim_msft_ad_unit_snapshot GENERATE
    $0 as ad_unit_id,
    $1 as ad_unit_name,
    $5 as account_id;

R3_source_term_detail = LOAD '/projects/apollo/source_term_detail/daily
/data/{20110101,20110102,20110103,20110104,20110105,20110106,20110107,20110108,20110109,20110110,20110111,20110112,2
0113,20110114,20110115,20110116,20110117,20110118}/part*' USING PigStorage('\u0001');
R4_source_term_detail = FOREACH R3_source_term_detail GENERATE
    $13 as ad_unit_id,
    $20 as bidded_searches,
    $22 as bidded_clicks,
    $41 as broad_revenue_pub_cur;

R5_output1 = JOIN R4_source_term_detail BY ad_unit_id,
    R2_dim_msft_ad_unit_snapshot BY ad_unit_id;

R6_output1 = FOREACH R5_output1 GENERATE
    $0 as source_term_detail_ad_unit_id,
    $1 as source_term_detail_bidded_searches,
    $2 as source_term_detail_bidded_clicks,
    $3 as source_term_detail_broad_revenue_pub_cur;

```



In My Opinion

- Is Cloud ready for DW & BI?
- Pros & Cons of BI on Cloud
- Options looked at:
 - Custom solution & Hive
 - Microstrategy & Hive
 - Pentaho



'Determine that things can and shall be done, and then we shall find the way!'

A. Lincoln