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

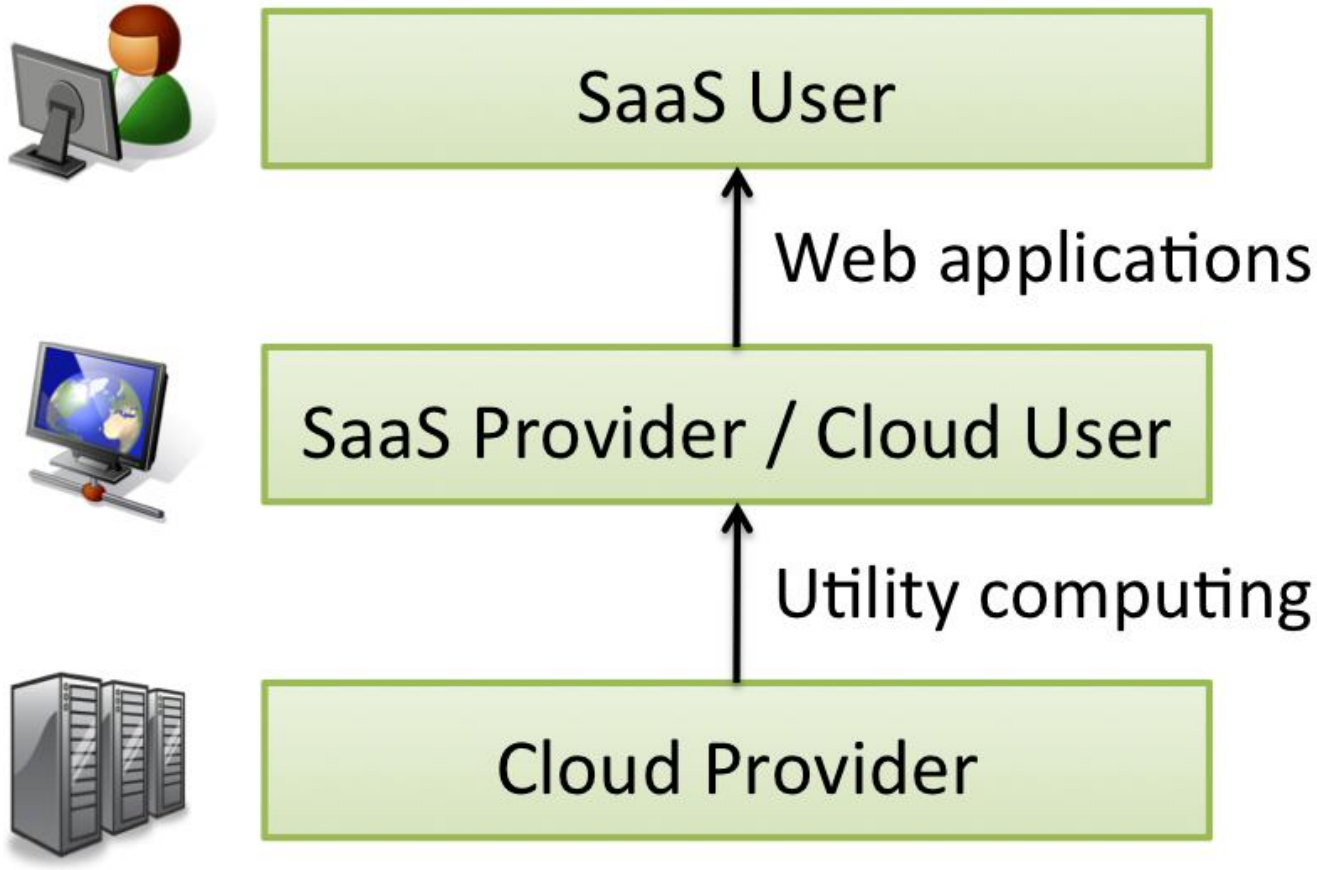
Cloud Computing

- 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



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 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

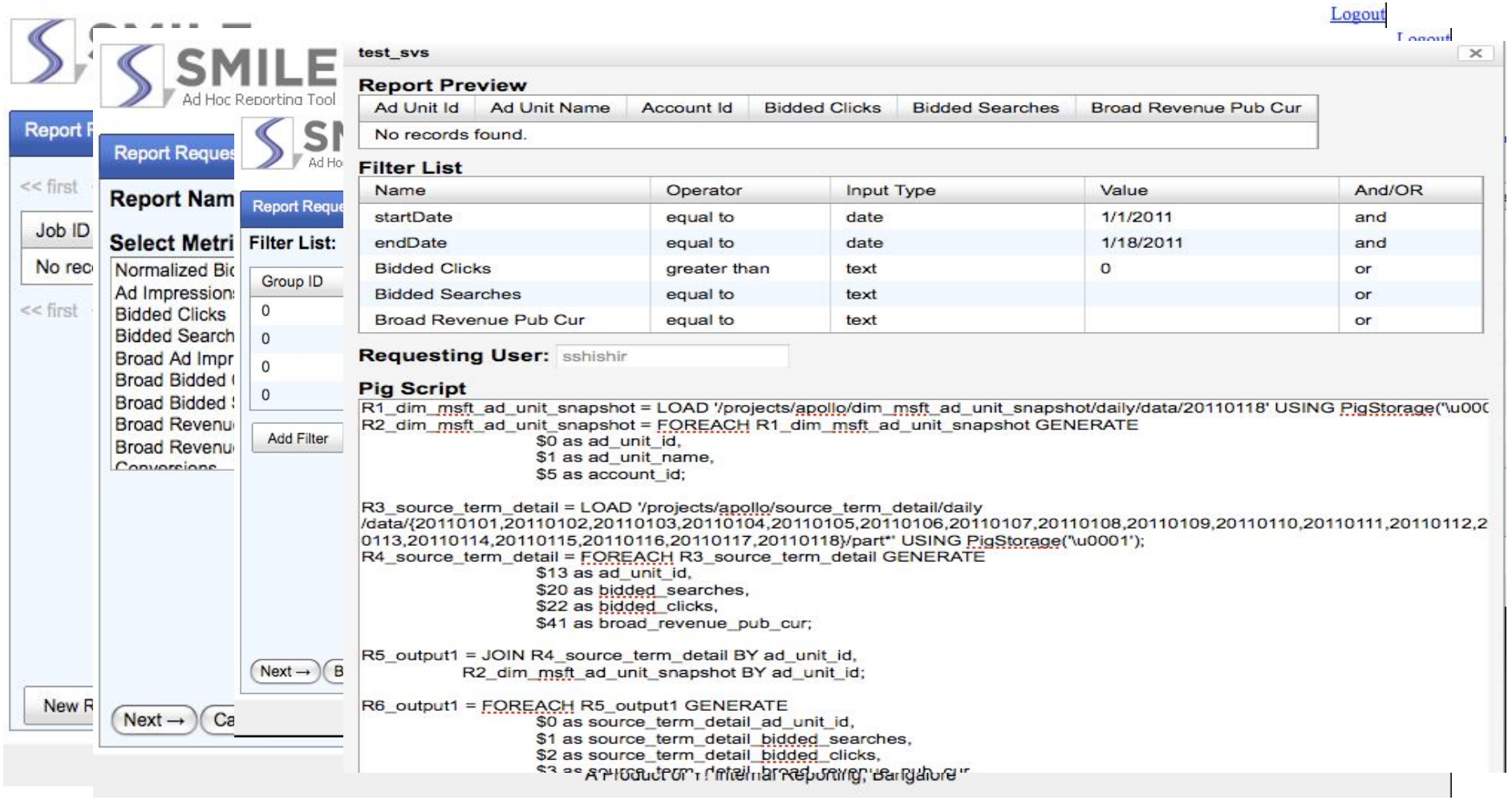
Rollups, Type 2 Dimension,

Alerts & Messaging

Metrics

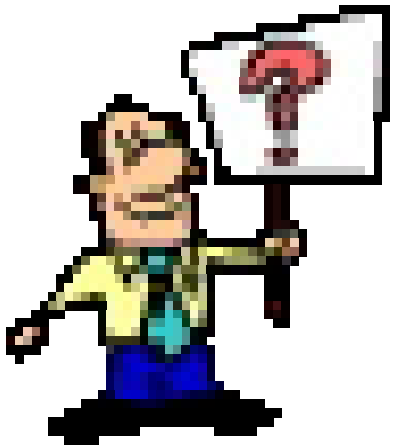
Impressions, Revenue, Clicks, Conversions, Quality Score, Top keywords

BI on Cloud – Screenshot



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