빅데이터 분석과 상용 클라우드의 결합.

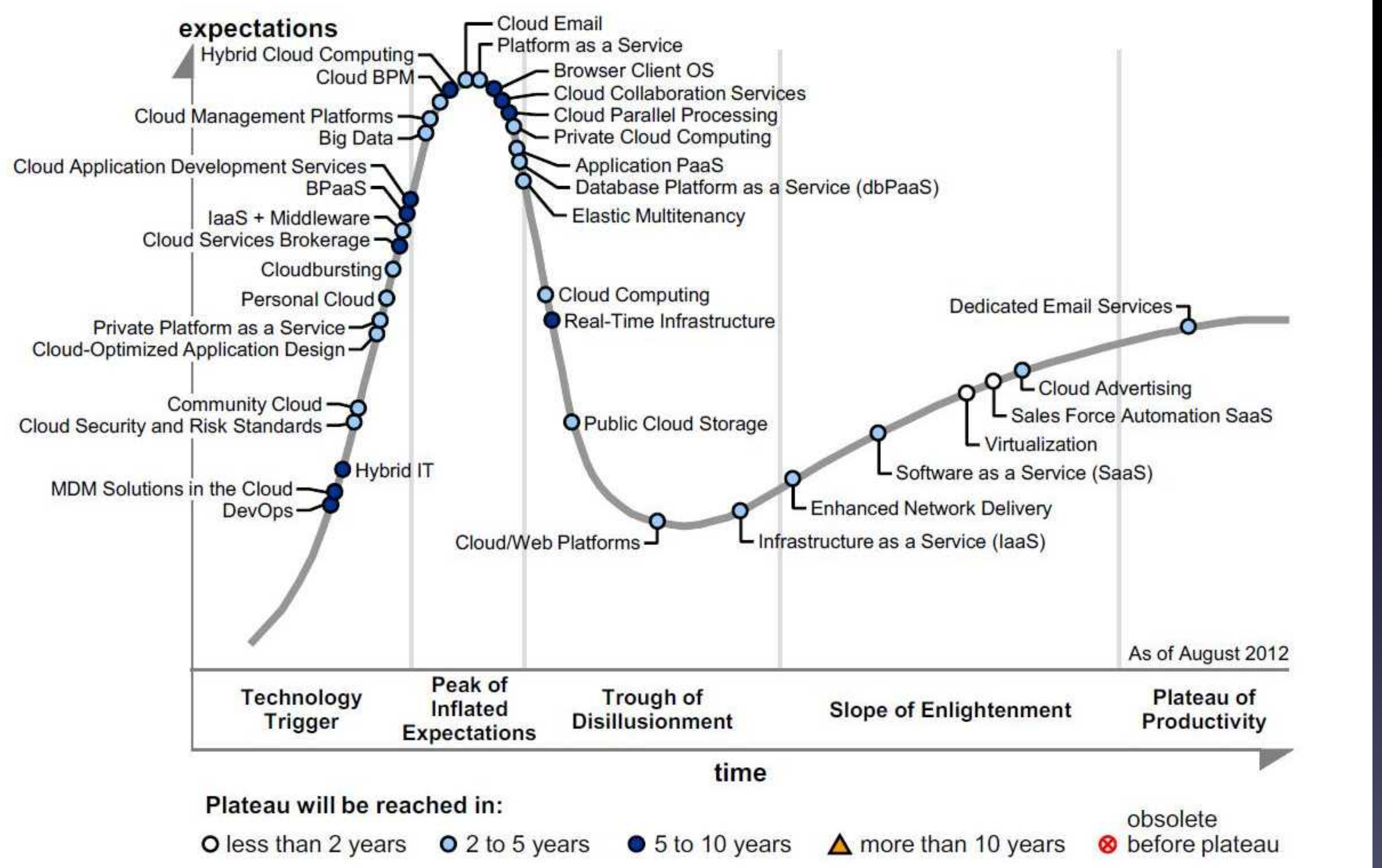
Big Data와 클라우드 컴퓨팅의 만남.

First Question?

- Why We try to connect the two huge word 'Big Data' & 'Cloud Computing'
  - – Can you define 'Big Data' , 'Cloud Computing'?
  - – isn't it just marketing words combining?

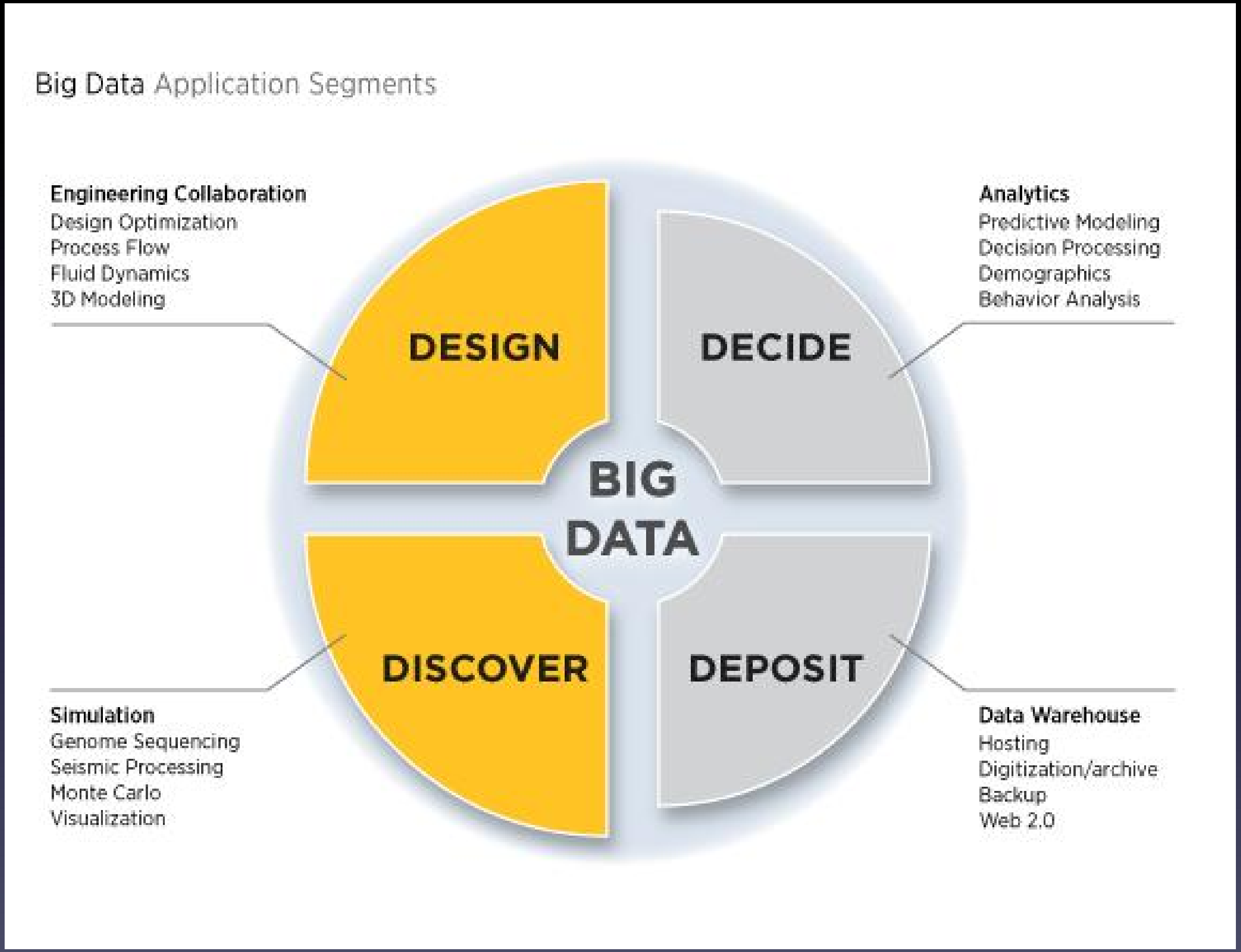
Big Data hype & Fever?

Figure 1. Hype Cycle for Cloud Computing, 2012



Source: Gartner (August 2012)

Big Data application segment



Big Data Application Segments

Engineering Collaboration
Design Optimization
Process Flow
Fluid Dynamics
3D Modeling

Analytics
Predictive Modeling
Decision Processing
Demographics
Behavior Analysis

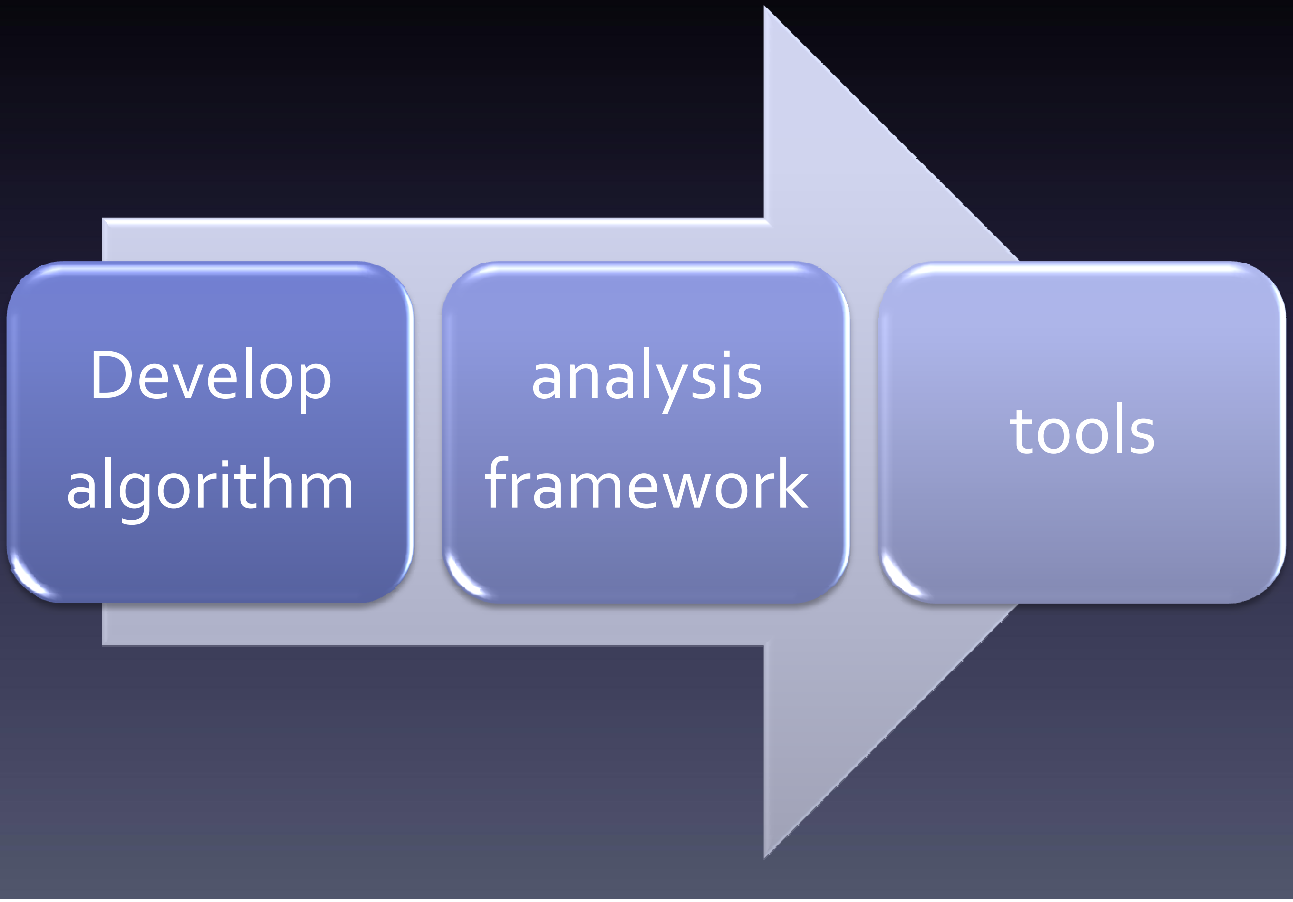
Simulation
Genome Sequencing
Seismic Processing
Monte Carlo
Visualization

Data Warehouse
Hosting
Digitization/archive
Backup
Web 2.0

source: panasas.com

Typical Decision for Data analysis.

- Things to decide when POC



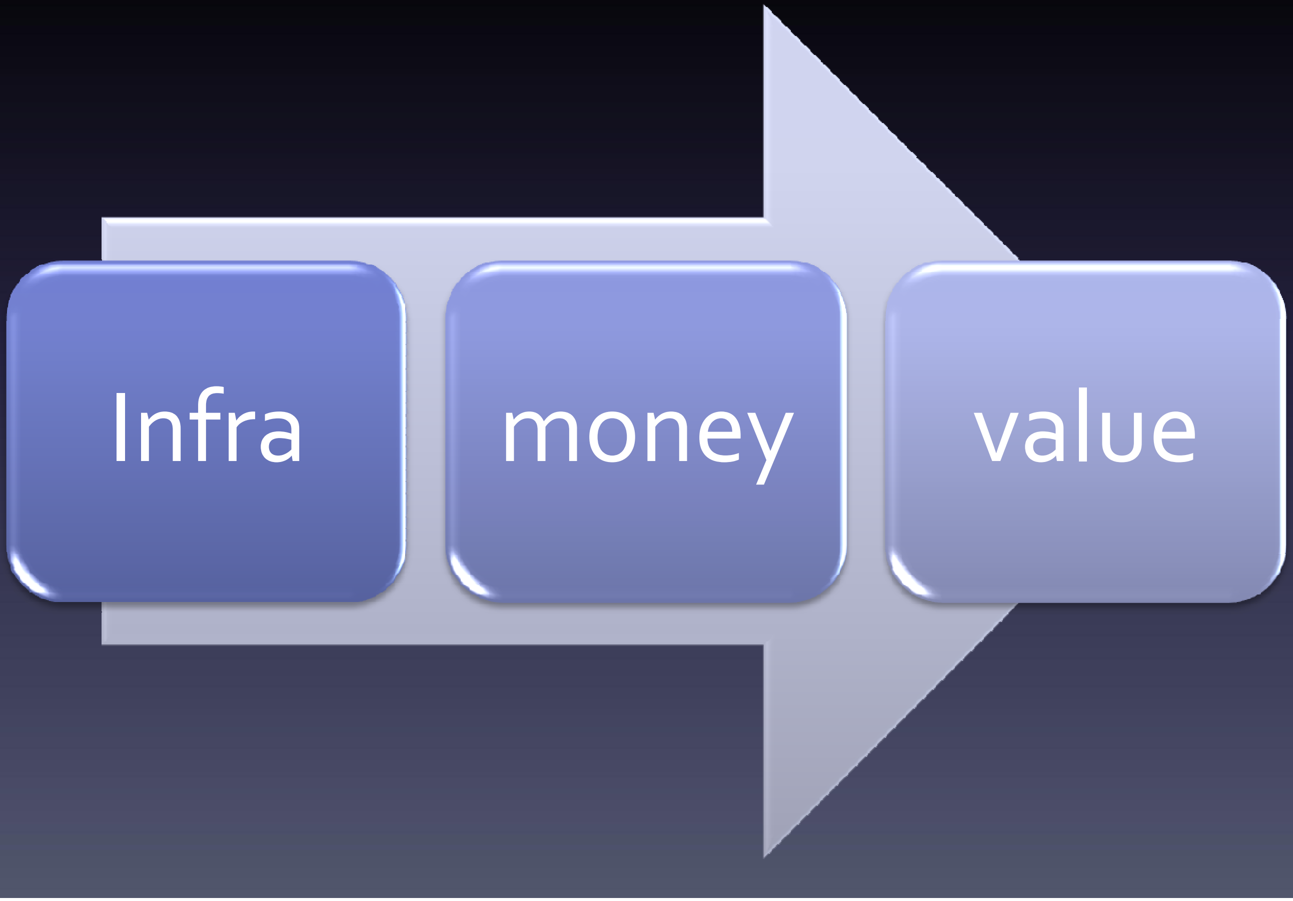
Develop algorithm

analysis framework

tools

Typical Decision for Data analysis.

- things to decide when goes out.

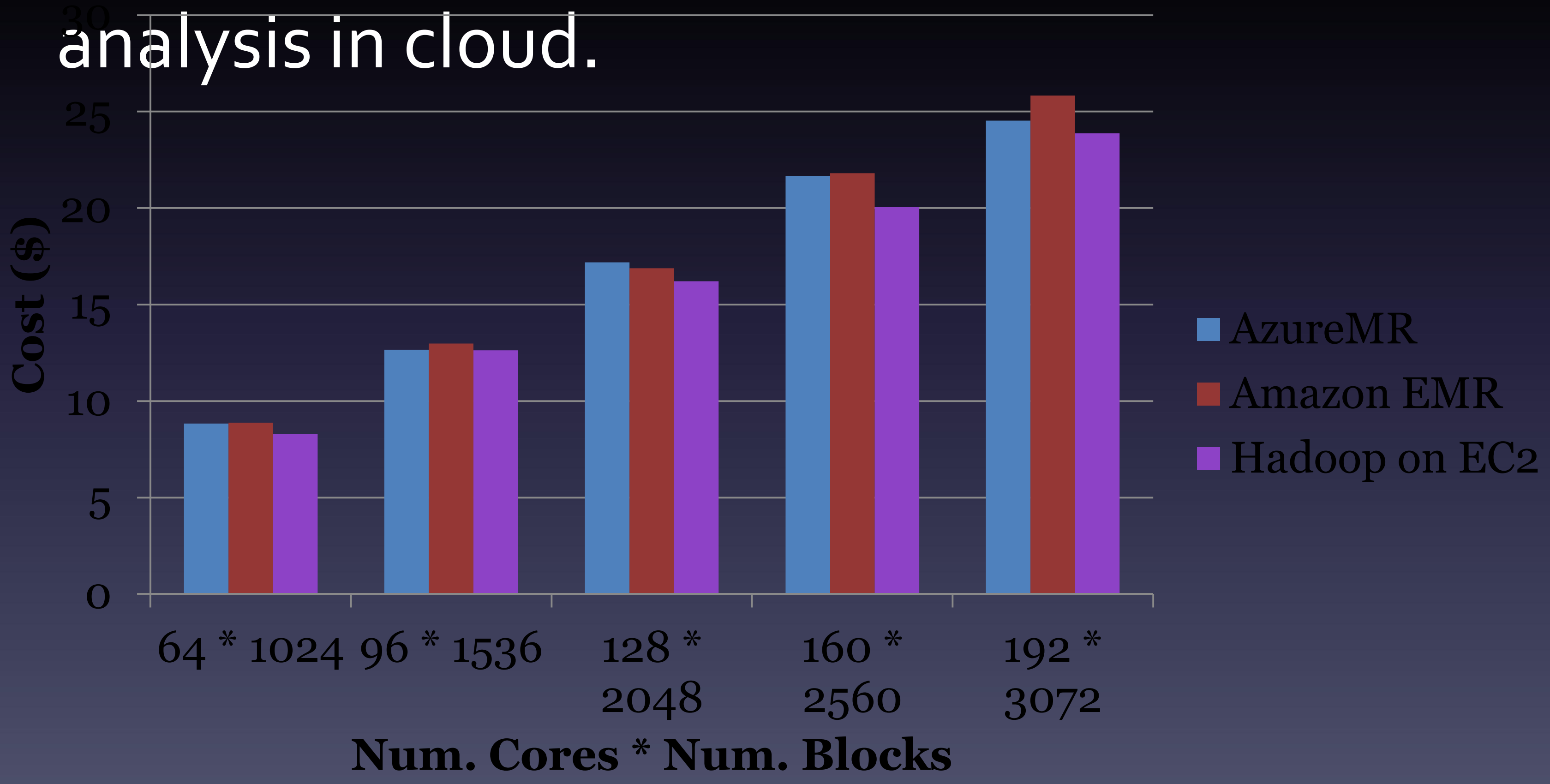


The example Case

- One Internet Fax company Case
  - they have 100,000 customer
  - the hosting cost of their machine is $24,000~30,000(including network cost)
  - Can you persuade the CEO of this company to invest a lot of money to build just 3 node hadoop cluster?

How much?

- How much When you develop multi-node analysis in cloud.



Time to implement

	Typical system	period
	Meeting -the server provider, network provider, IDC manager	1month
	Order - Place order, you check importing schedule(always delayed)	1month
	Building -installation Rack, Server, Service slide, networking (always missing/broken parts)	1month
	Testing -check OS , Network, Disk -install your application	1month

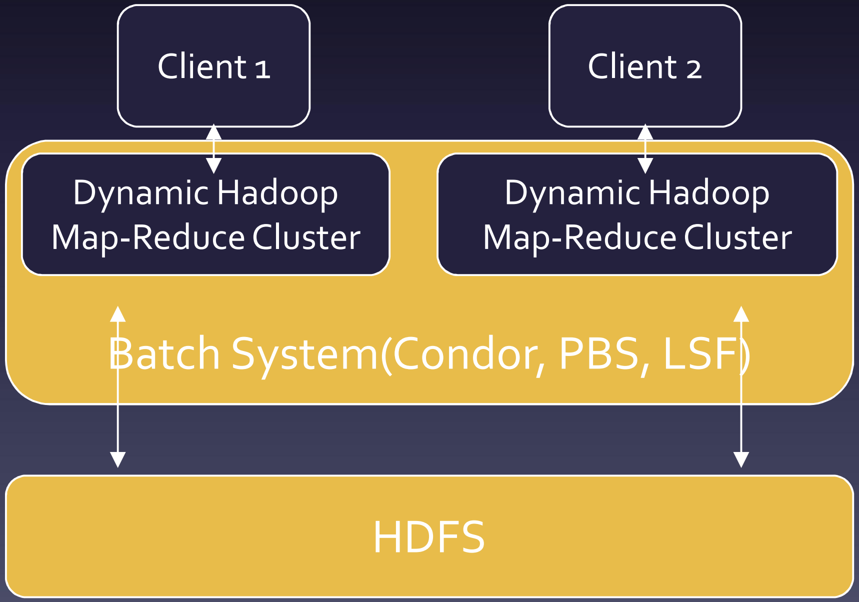
	Cloud infra	period
	Check the web site -punch in your credit card number	1hour
	Order - Make 256 virtual instance	1~2 day
	Building - Waiting all instances coming up	1~2 days
	Testing -install your application -run your test	1 month
	Done - Erase the instance	2~4 hour

computing framework to cloud

- MapReduce Framework (Bare-Metal Style)
  - Hadoop on EC2
  - Hadoop on google compute.
  - Hadoop on Azure instance.

computing framework to cloud

- Pre-Configured Hadoop
  - the 1st generation: HOD(Hadoop On Demand)


Client 1
Client 2
Dynamic Hadoop Map-Reduce Cluster
Dynamic Hadoop Map-Reduce Cluster
Batch System(Condor, PBS, LSF)
HDFS

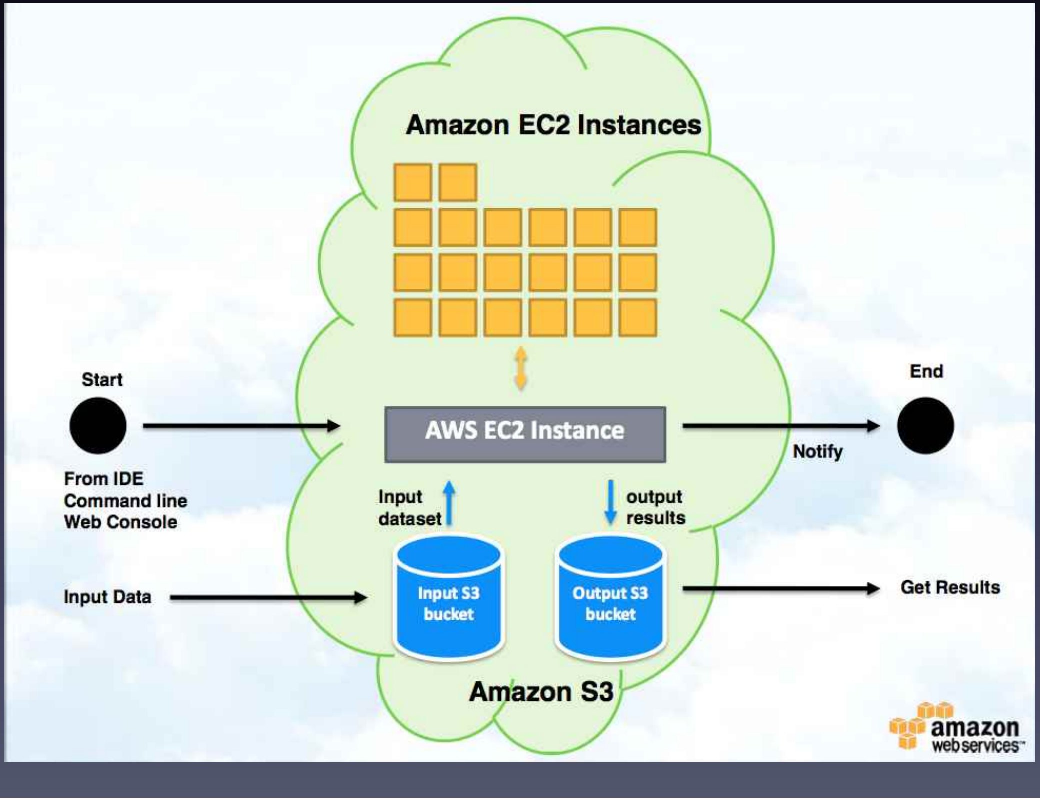
Script Sample
hod allocate -d cluster_dir –n 16
hadoop --config ~/hod-clusters/test jar /hadoop-examples.jar wordcount /input /path/to/output

computing framework to cloud

- Pre-Configured Hadoop
  - the 2nd generation: EMR(Elastic MapReduce)

Amazon Elastic MapReduce =

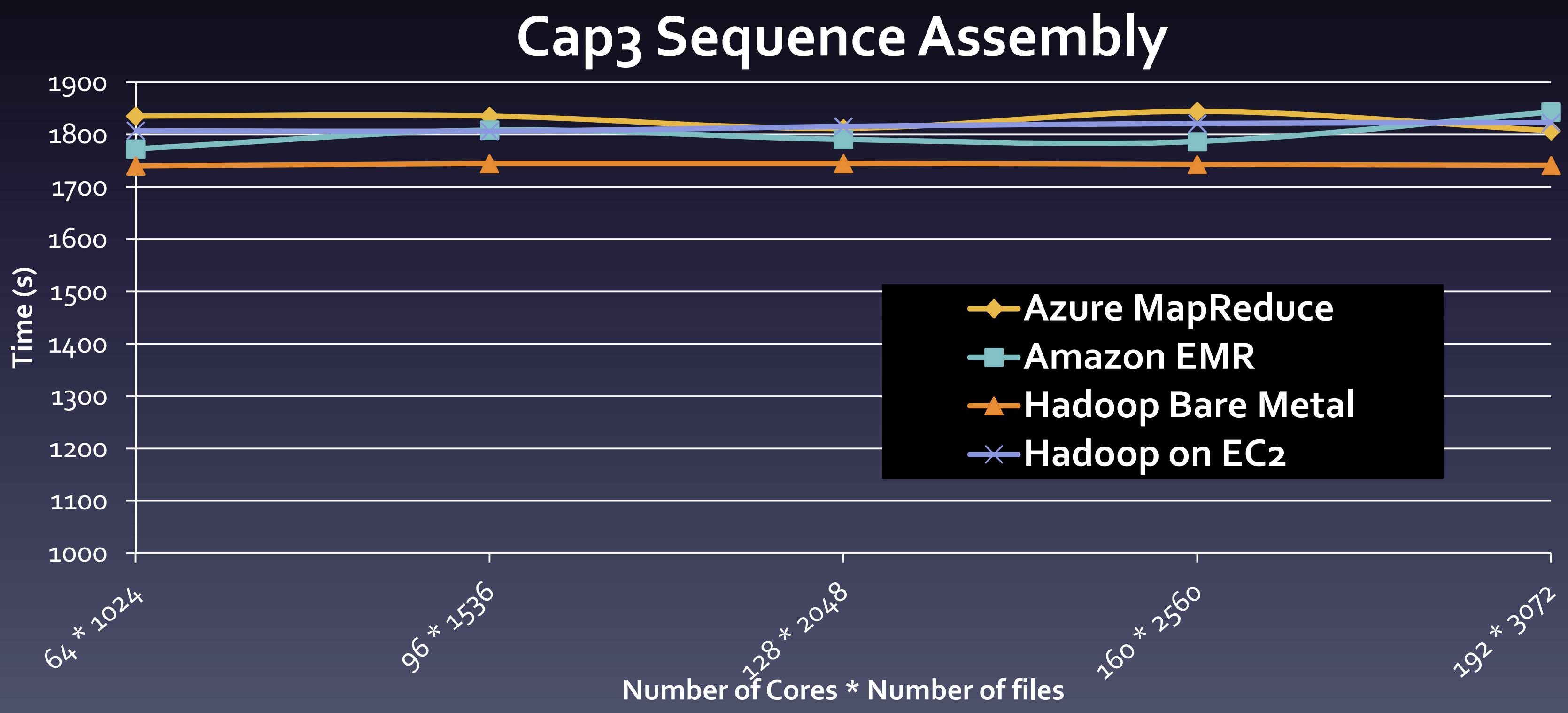
Amazon EC2 + Hadoop



source: AWS

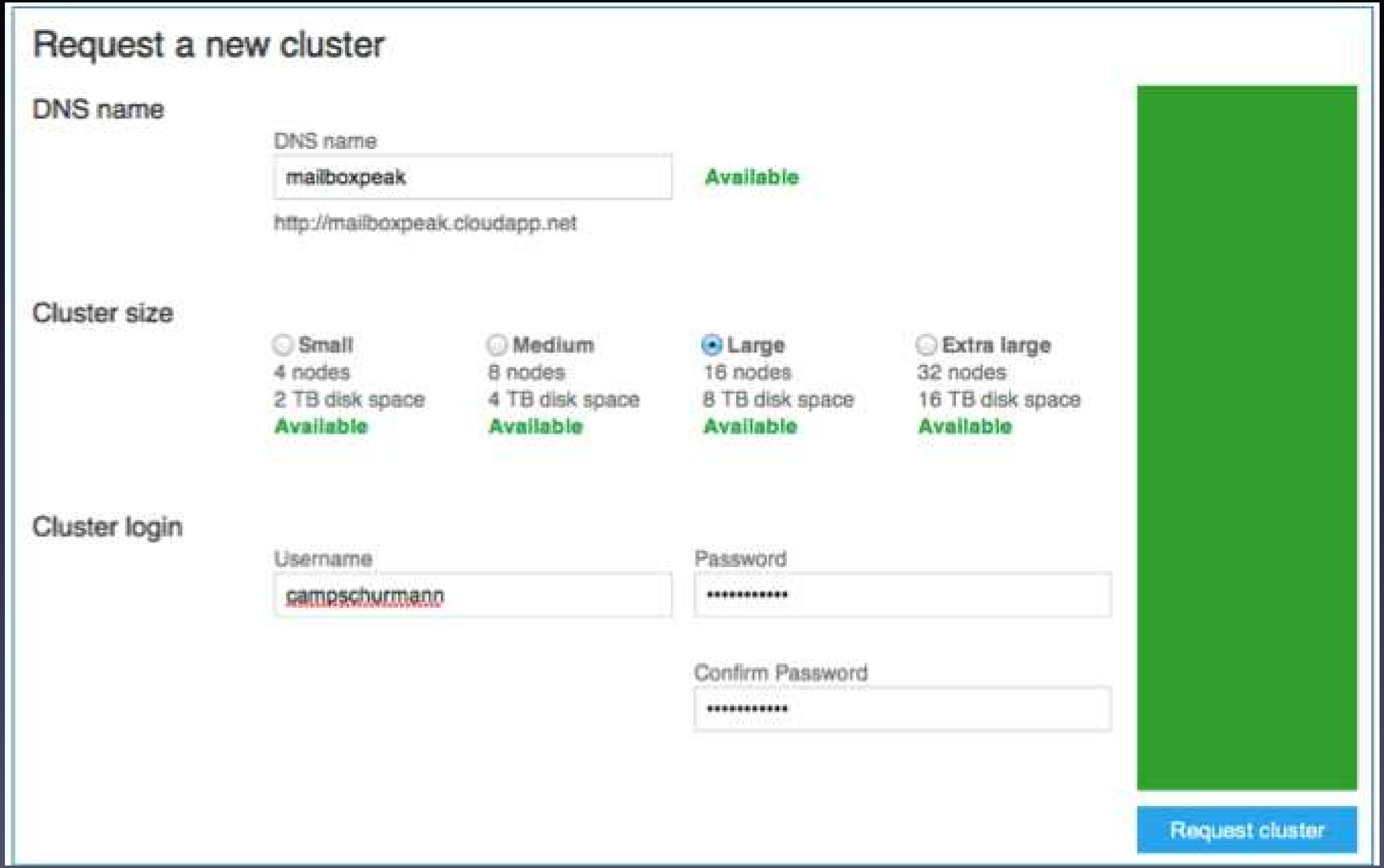
computing framework to cloud

- Pre-Configured Hadoop
  - the 2nd generation: EMR(Elastic MapReduce)



computing framework to cloud

- Pre-configured hadoop – hadoopOnAzure



Request a new cluster

DNS name

DNS name: mailboxpeak — Available

http://mailboxpeak.cloudapp.net

Cluster size

○ Small	○ Medium	● Large	○ Extra large
4 nodes	8 nodes	16 nodes	32 nodes
2 TB disk space	4 TB disk space	8 TB disk space	16 TB disk space
Available	Available	Available	Available

Cluster login

Username: campschurmann

Password: ••••••••••••

Confirm Password: ••••••••••••

Computing framework to cloud

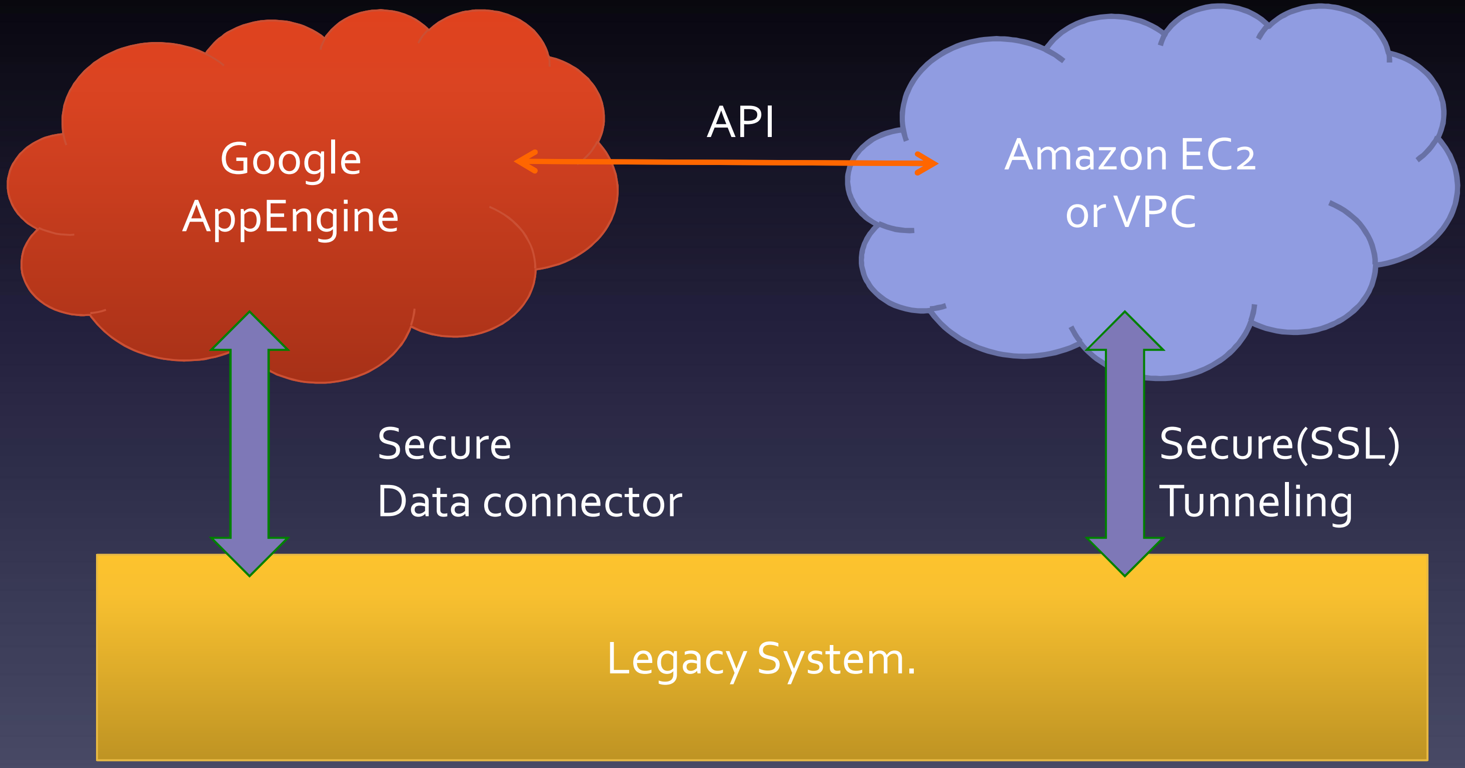
- EMR - Demo

Computing framework to cloud

- Cloud only for MapReduce?
  - What about MPI(Message Passing Interface) based Clusters?
  - Most of Commercial Parallel Solutions developed in MPI library.

Computing framework to cloud

- Virtual Private Cloud

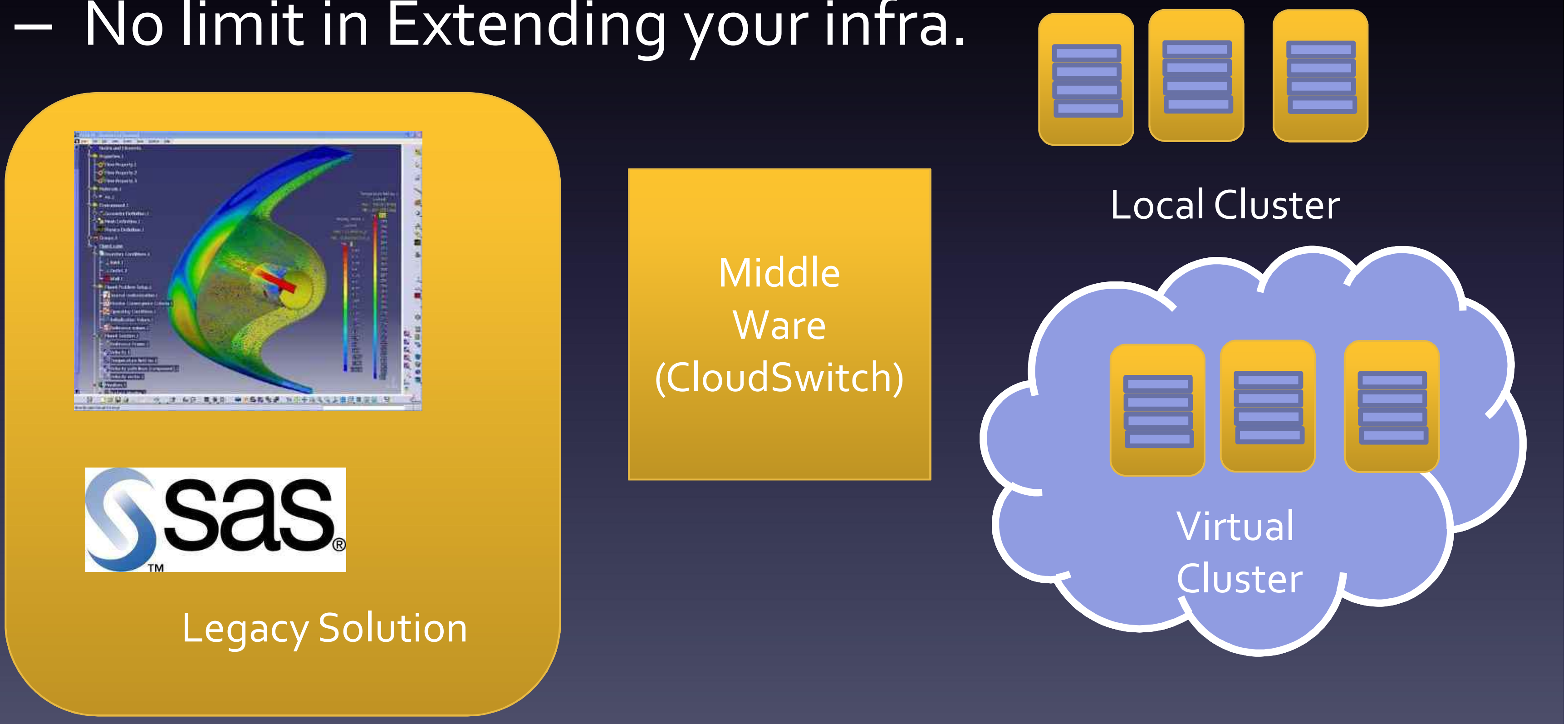


Computing framework to cloud

- Cloudburst of Legacy System.

Computing framework to cloud

- Cloudburst of Legacy System.
  - No limit in Extending your infra.



Legacy Solution

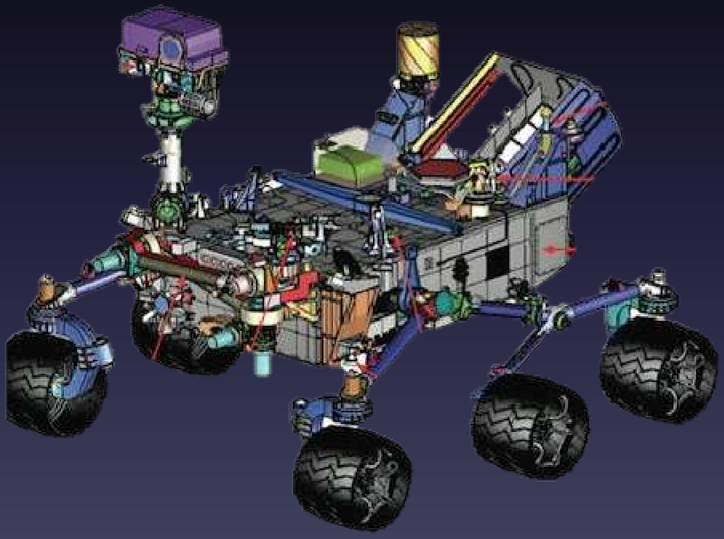
Middle Ware (CloudSwitch)

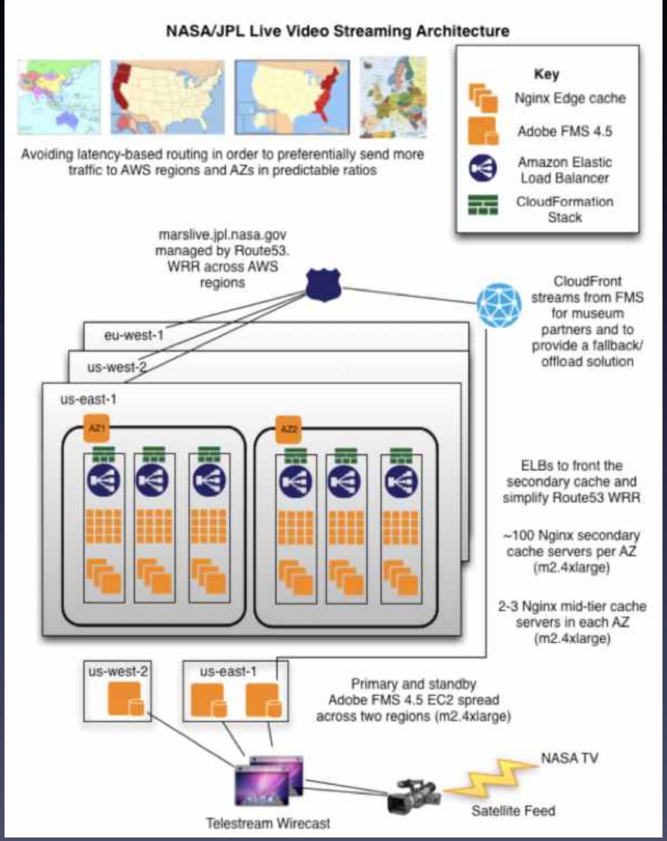
Local Cluster

Virtual Cluster

Computing framework to Cloud

- Do you Remember the "Curiosity?"





NASA/JPL Live Video Streaming Architecture

Avoiding latency-based routing in order to preferentially send more traffic to AWS regions and AZs in predictable ratios

Key

- Nginx Edge cache
- Adobe FMS 4.5
- Amazon Elastic Load Balancer
- CloudFormation Stack

marslive.jpl.nasa.gov managed by Route53. WRR across AWS regions

CloudFront streams from FMS for museum partners and to provide a fallback/ offload solution

eu-west-1

us-west-2

us-east-1

AZ1

AZ2

ELBs to front the secondary cache and simplify Route53 WRR

~100 Nginx secondary cache servers per AZ (m2.4xlarge)

2-3 Nginx mid-tier cache servers in each AZ (m2.4xlarge)

us-west-2

us-east-1

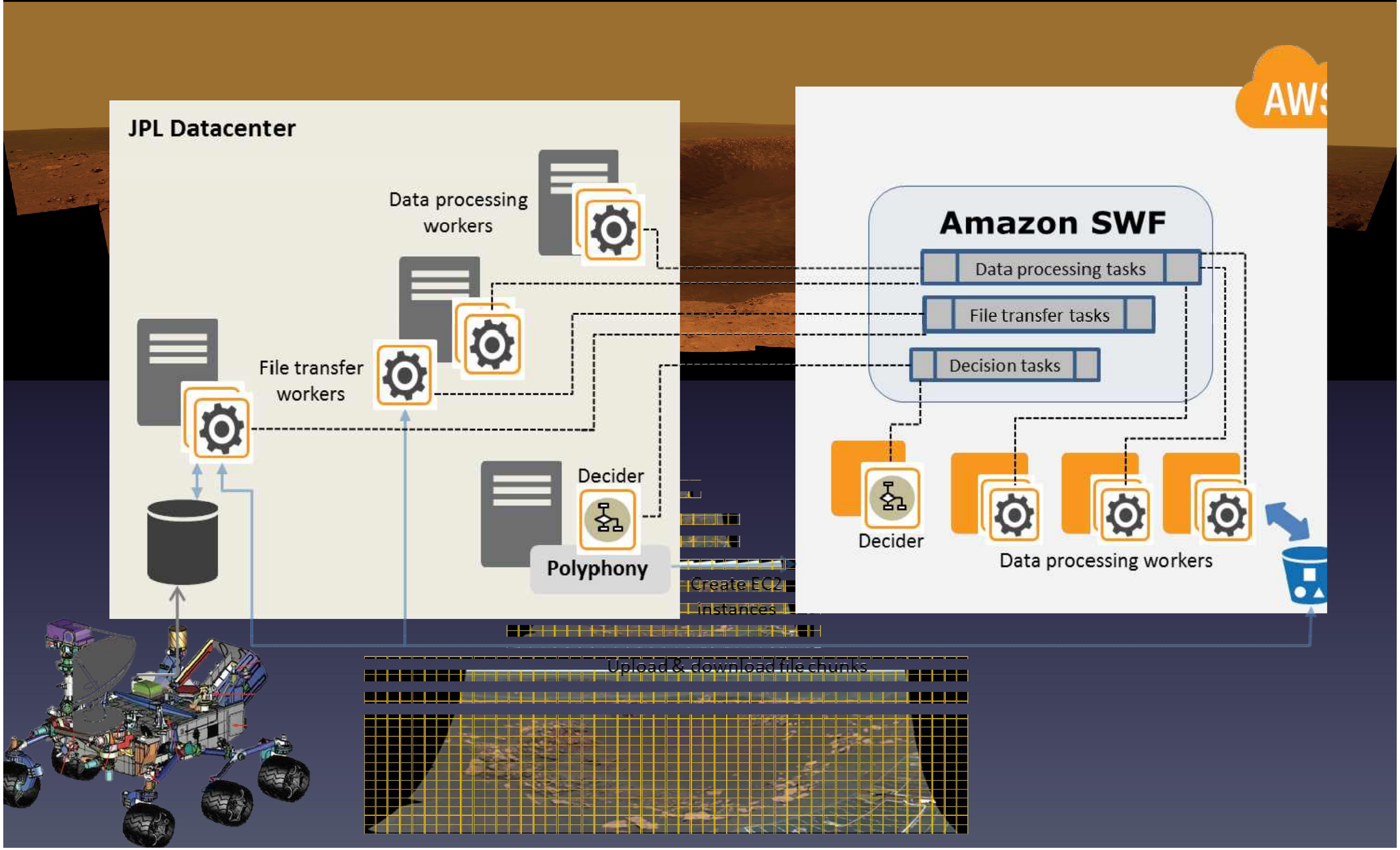
Primary and standby Adobe FMS 4.5 EC2 spread across two regions (m2.4xlarge)

NASA TV

Satellite Feed

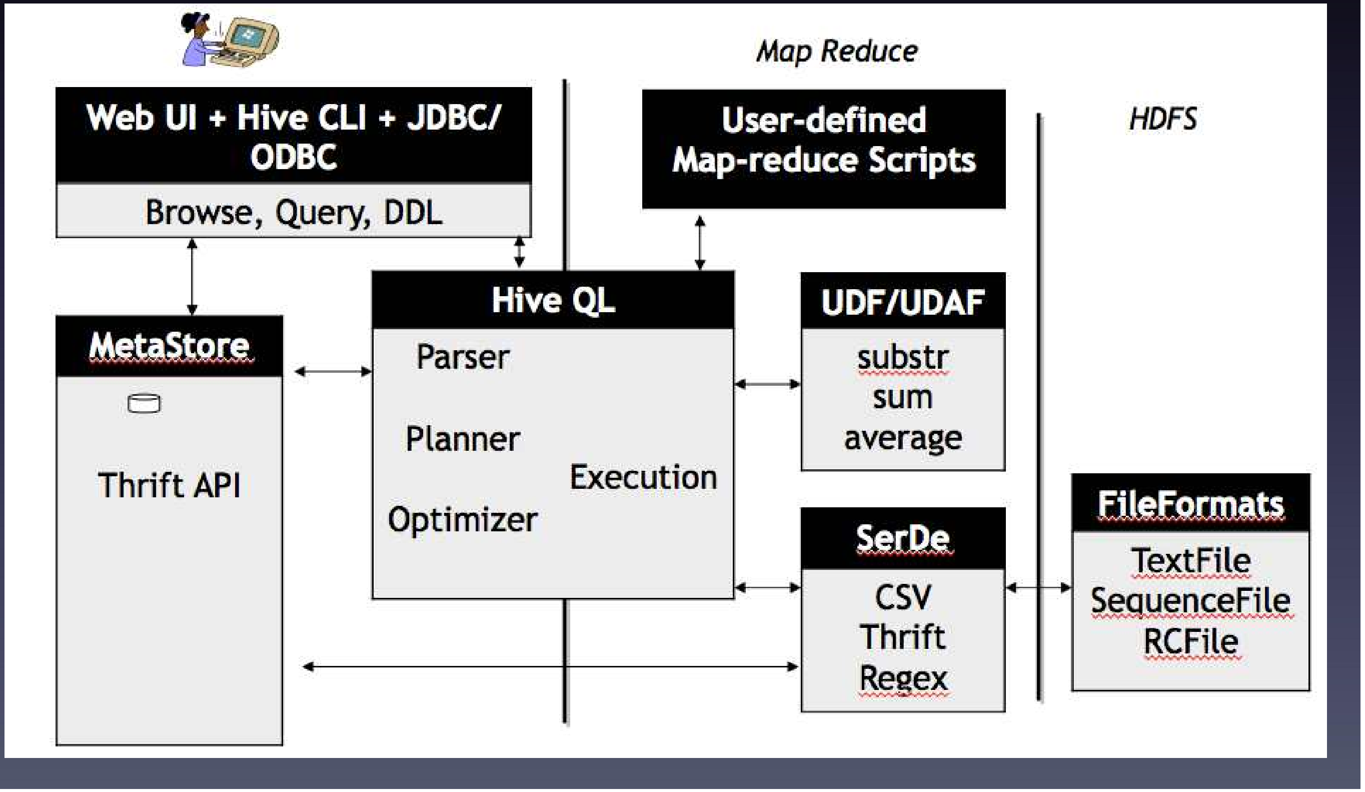
Telestream Wirecast

Computing framework to Cloud



Computing framework to cloud

- Query based Analytic Engine
  - Hive(hive.apache.org)



Map Reduce

HDFS

Web UI + Hive CLI + JDBC/ ODBC

Browse, Query, DDL

User-defined Map-reduce Scripts

Hive QL

Parser

Planner

Execution

Optimizer

UDF/UDAF

substr
sum
average

MetaStore

Thrift API

SerDe

CSV
Thrift
Regex

FileFormats

TextFile
SequenceFile
RCFile

Computing framework to cloud

- Hive on Cloud
  - Hive Job In EMR
  - Interactive mode

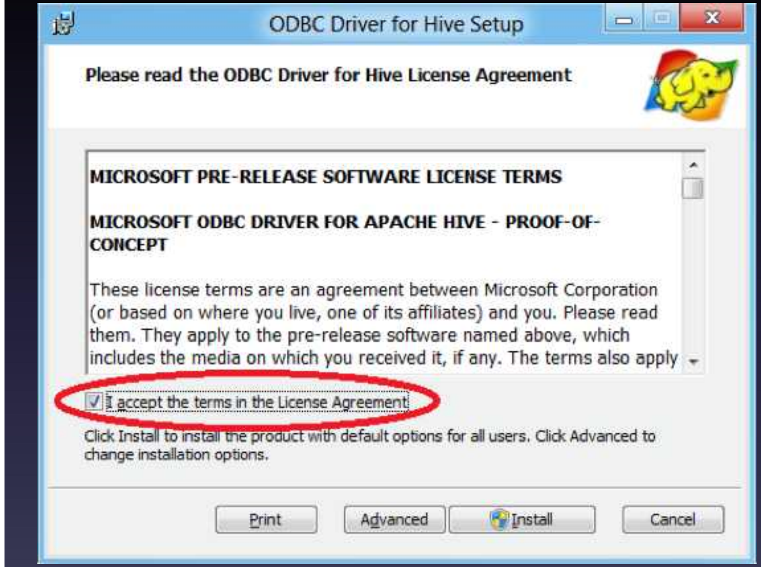
```
./elastic-mapreduce --create --name
"${JOB_NAME}"
   --hive-interactive --num-instances
${EMR_INSTANCES_NUM}
   --master-instance
${EMR_INSTANCES_TYPE} --alive
```

  - script mode

```
./elastic-mapreduce --create \
   --hive-script --args ${EMR_SCRIPT_PATH} \
   --args -
d,OUTPUT_PATH=${OUTPUT_LOCATION_S3} \
   --name "${JOB_NAME}" \
   --num-instances ${EMR_INSTANCES_NUM} \
   --instance-type ${EMR_INSTANCES_TYPE} \
   --credentials ${EMR_CREDENTIALS_FILE})
```

BI(Business Intelligence) with HIVE

- EXCEL(most popular BI)



Please read the ODBC Driver for Hive License Agreement

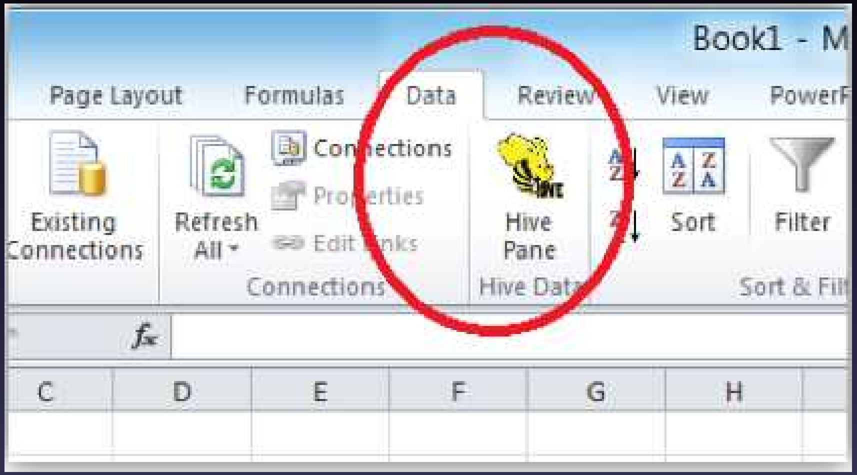
MICROSOFT PRE-RELEASE SOFTWARE LICENSE TERMS

MICROSOFT ODBC DRIVER FOR APACHE HIVE - PROOF-OF-CONCEPT

These license terms are an agreement between Microsoft Corporation (or based on where you live, one of its affiliates) and you. Please read them. They apply to the pre-release software named above, which includes the media on which you received it, if any. The terms also apply

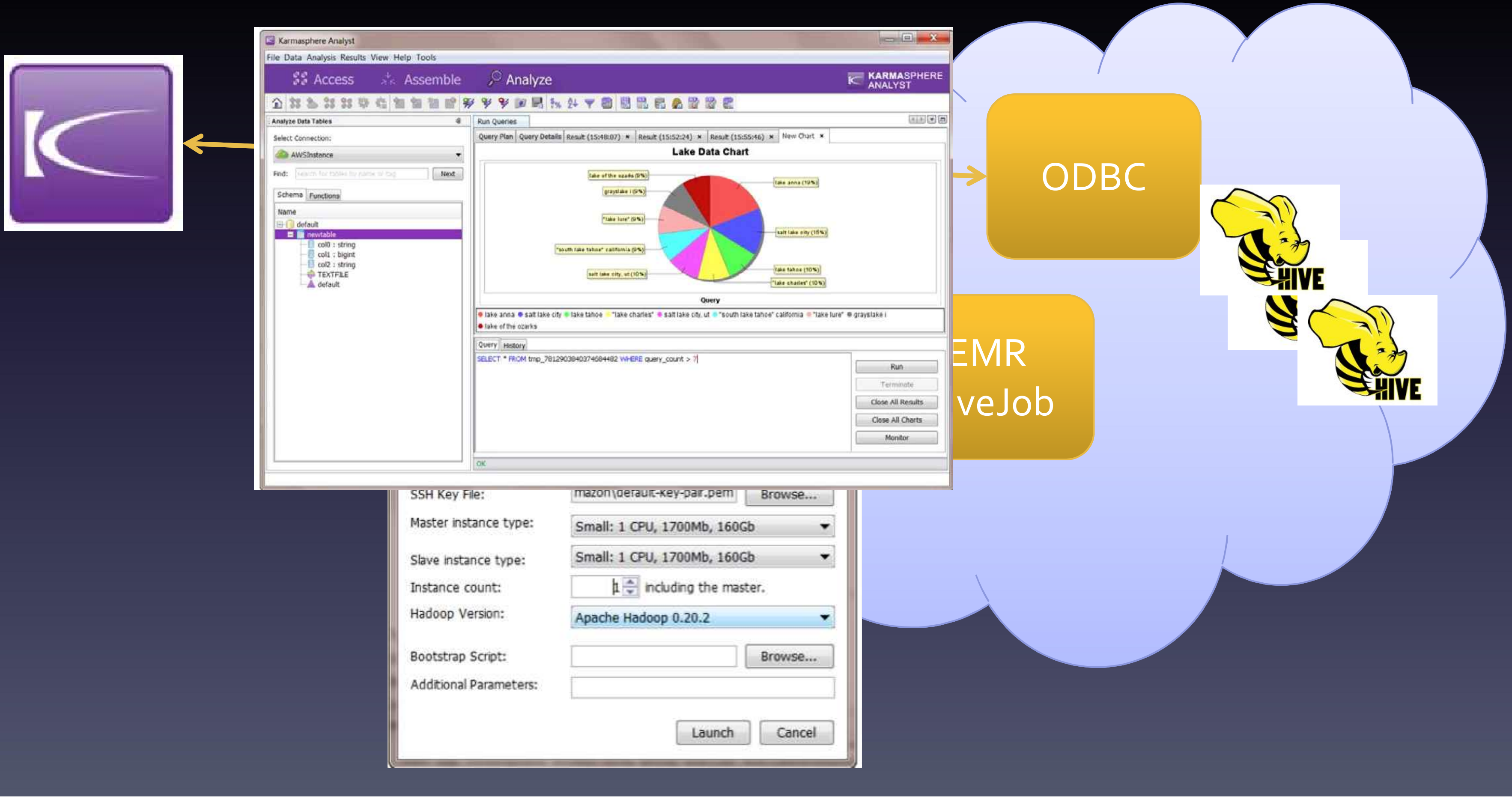
☑ I accept the terms in the License Agreement

Click Install to install the product with default options for all users. Click Advanced to change installation options.



BI(Business Intelligence) with HIVE

- Karmasphere BI



Select Connection: AWSInstance

Find: search for tables by name or tag [Next]

```
SELECT * FROM tmp_7812903840374684482 WHERE query_count > 7
```

SSH Key File: mazon\default-key-pair.pem

Master instance type: Small: 1 CPU, 1700Mb, 160Gb

Slave instance type: Small: 1 CPU, 1700Mb, 160Gb

Instance count: 1 including the master.

Hadoop Version: Apache Hadoop 0.20.2

Bootstrap Script:

Additional Parameters: