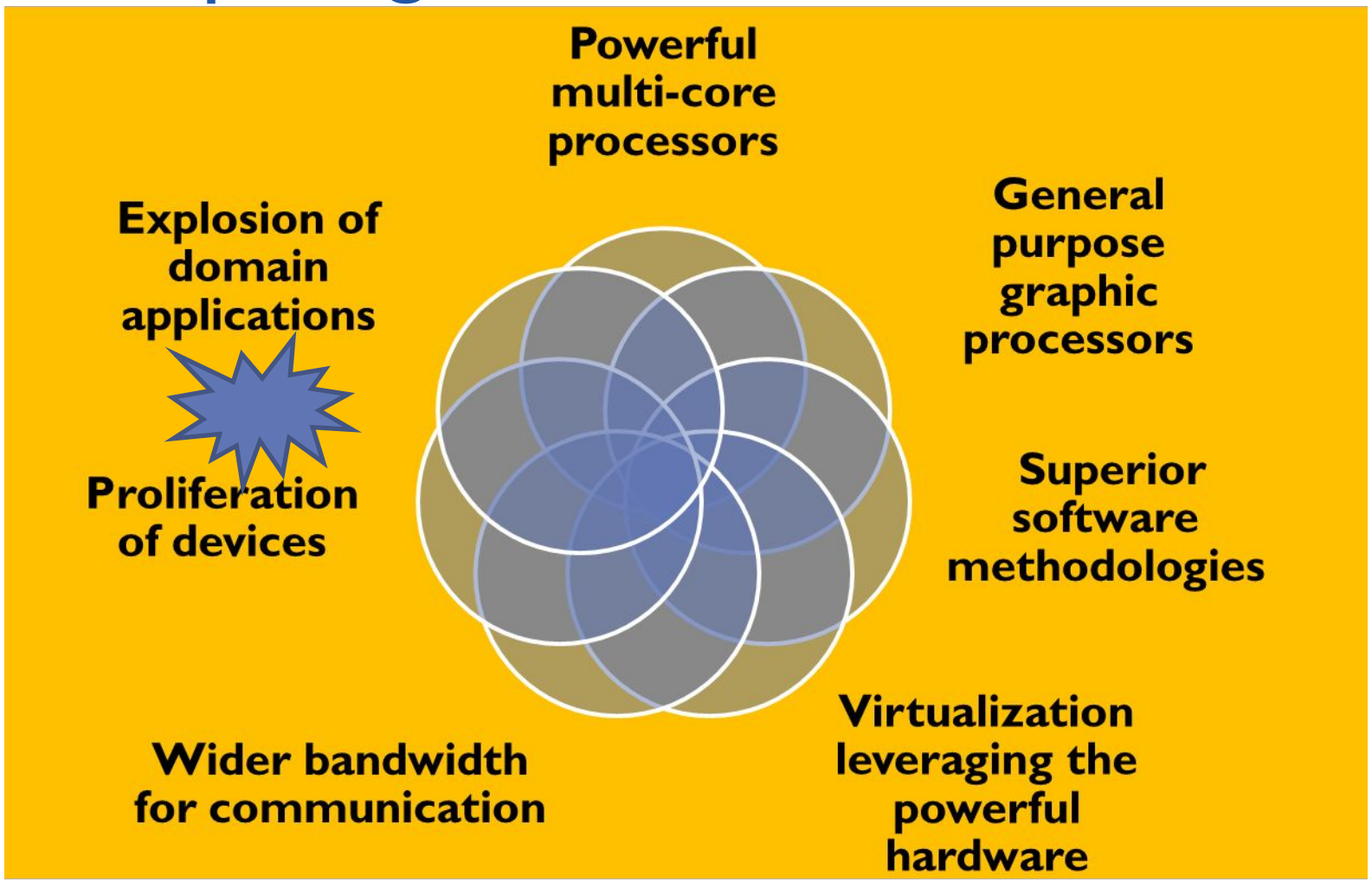
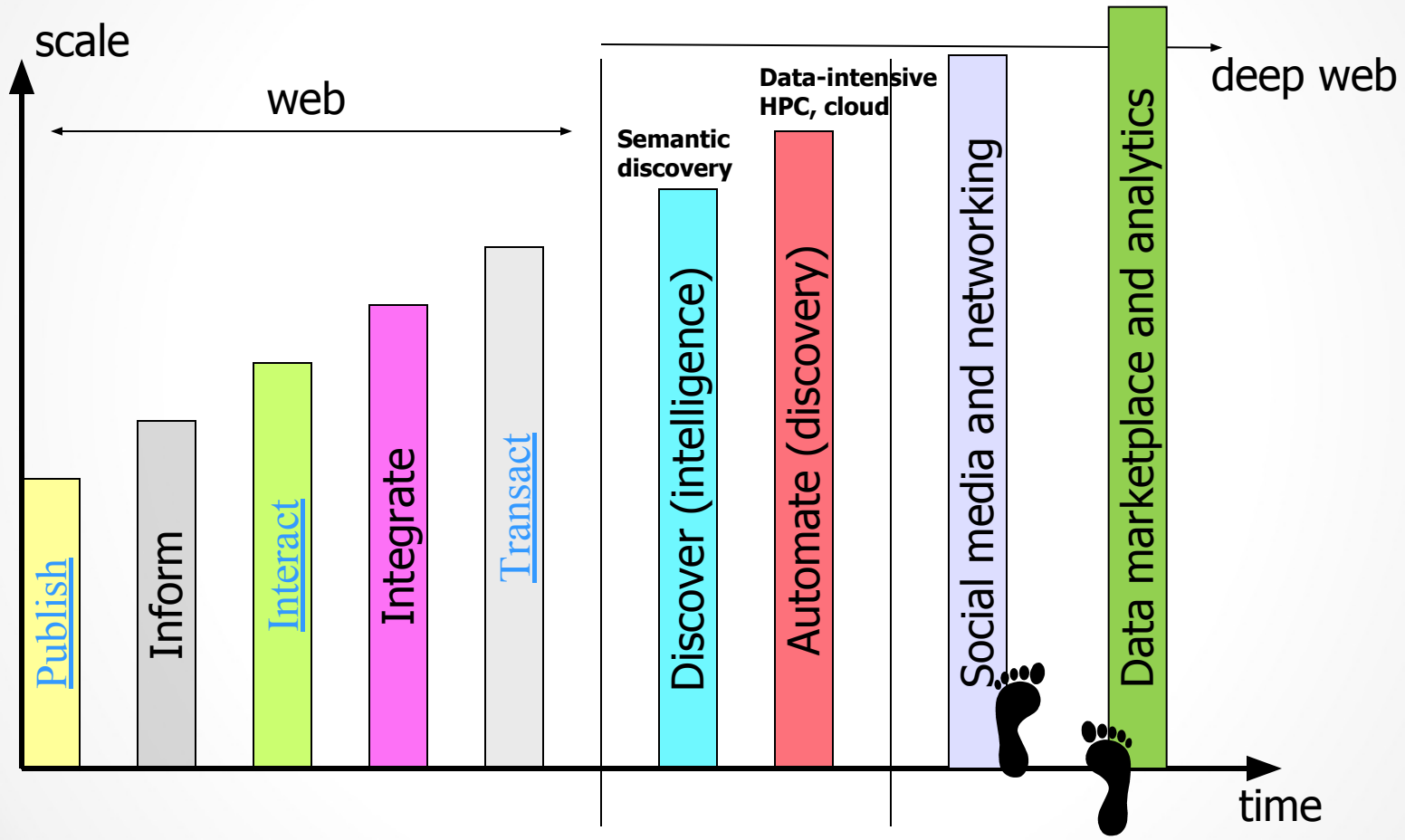
Cloud Computing: Concepts, Technologies and Business Implications

Introduction: A Golden Era in Computing

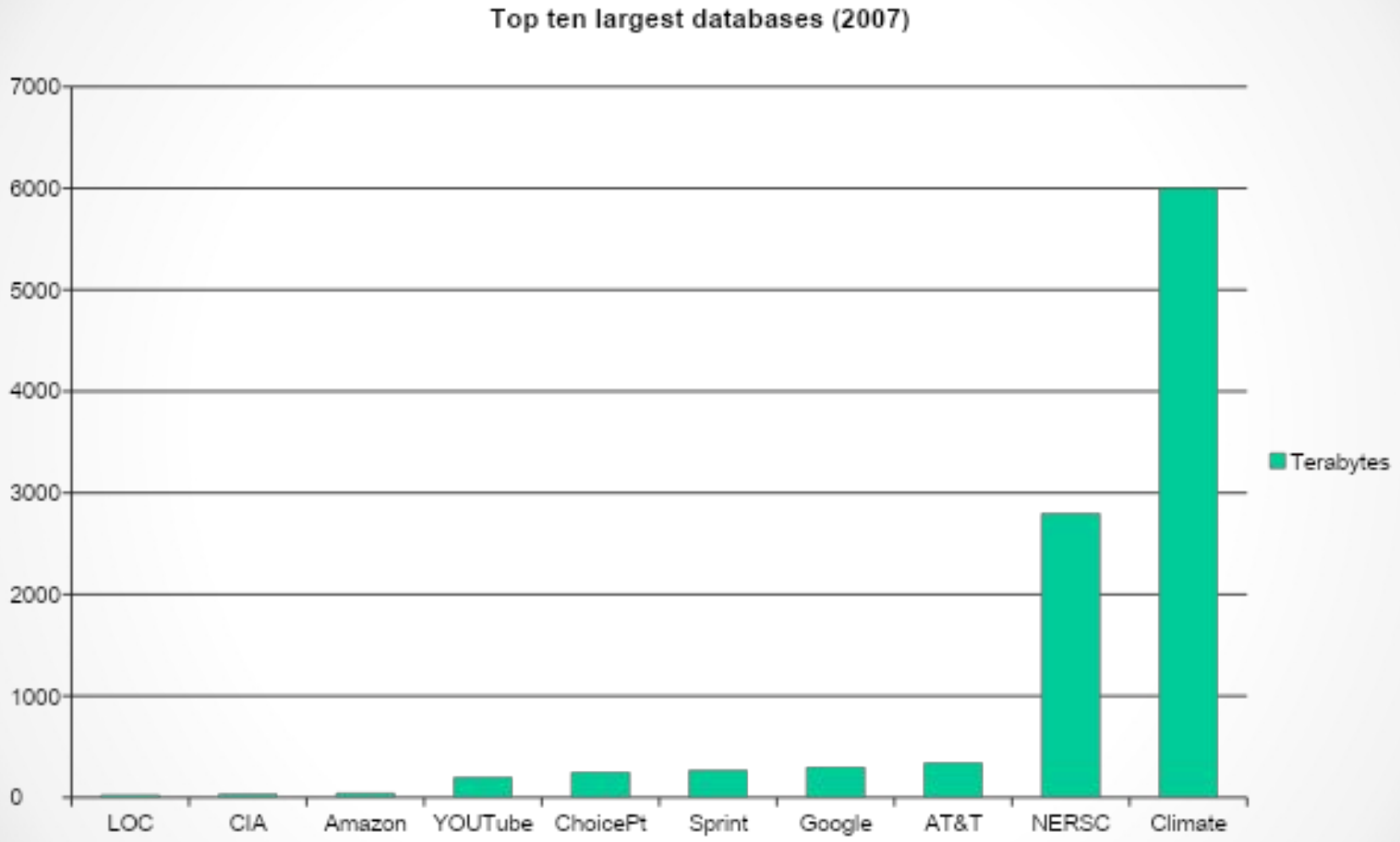


Cloud Concepts, Enabling-technologies, and Models: The Cloud Context

Evolution of Internet Computing



Top Ten Largest Databases



Ref: http://www.focus.com/fyi/operations/10-largest-databases-in-the-world/

Challenges

- Alignment with the needs of the business / user / non-computer specialists / community and society
- Need to address the scalability issue: large scale data, high performance computing, automation, response time, rapid prototyping, and rapid time to production
- Need to effectively address (i) ever shortening cycle of obsolescence, (ii) heterogeneity and (iii) rapid changes in requirements
- Transform data from diverse sources into intelligence and deliver intelligence to right people/user/systems
- What about providing all this in a cost-effective manner?

Enter the cloud

- **Cloud computing** is Internet-based computing, whereby shared resources, software and information are provided to computers and other devices on-demand, like the electricity grid.
- The cloud computing is a culmination of numerous attempts at large scale computing with seamless access to virtually limitless resources.
  - on-demand computing, utility computing, ubiquitous computing, autonomic computing, platform computing, edge computing, elastic computing, **grid computing**, …

"Grid Technology"

- Emerging enabling technology.
- Natural evolution of distributed systems and the Internet.
- Middleware supporting network of systems to facilitate sharing, standardization and openness.
- Infrastructure and application model dealing with sharing of compute cycles, data, storage and other resources.
- Publicized by prominent industries as on-demand computing, utility computing, etc.
- Move towards delivering "computing" to masses similar to other utilities (electricity and voice communication)."
- Now,

Hmmm…sounds like the definition for cloud computing!!!!!

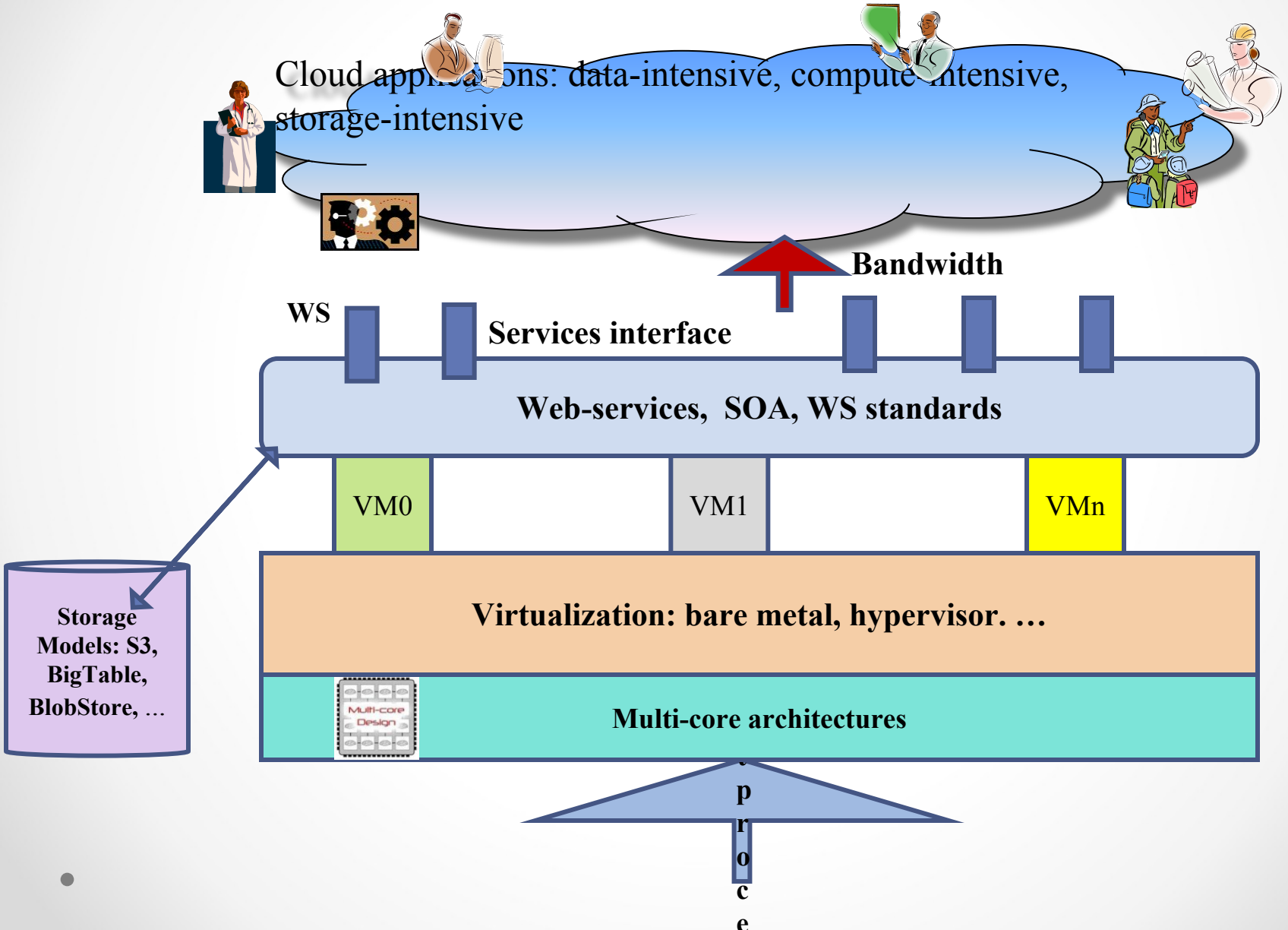
It is a changed world now…

- Explosive growth in applications: biomedical informatics, space exploration, business analytics, web 2.0 social networking: YouTube, Facebook
- Extreme scale content generation: e-science and e-business data deluge
- Extraordinary rate of digital content consumption: digital gluttony: Apple iPhone, iPad, Amazon Kindle
- Exponential growth in compute capabilities: multi-core, storage, bandwidth, virtual machines (virtualization)
- Very short cycle of obsolescence in technologies: Windows Vista Windows 7; Java versions; C C#; Phython
- Newer architectures: web services, persistence models, distributed file systems/repositories (Google, Hadoop), multi-core, wireless and mobile
- Diverse knowledge and skill levels of the workforce
- You simply cannot manage this complex situation with your traditional IT infrastructure:

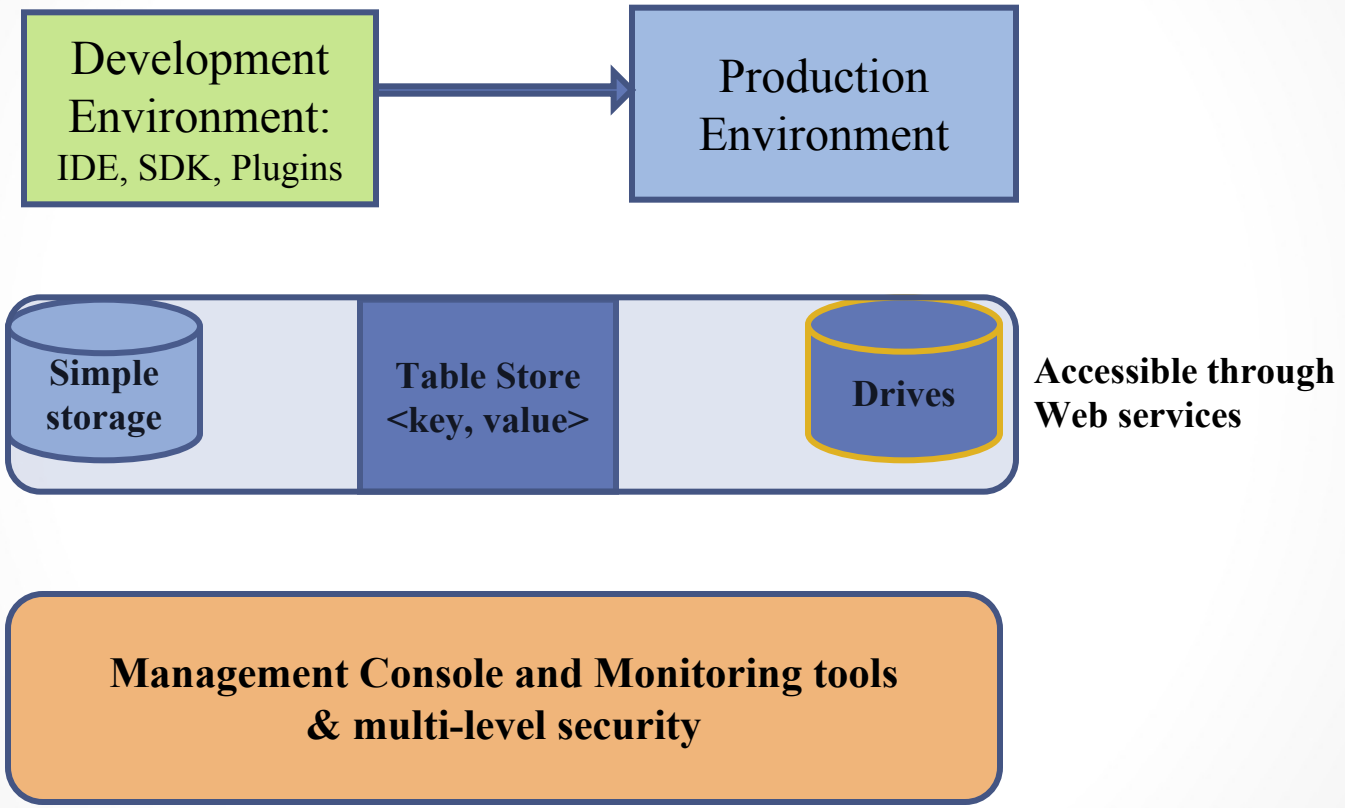
Answer: The Cloud Computing?

- Typical requirements and models:
  - platform (PaaS),
  - software (SaaS),
  - infrastructure (IaaS),
  - Services-based application programming interface (API)
- A cloud computing environment can provide one or more of these requirements for a cost
- Pay as you go model of business
- When using a public cloud the model is similar to renting a property than owning one.
- An organization could also maintain a private cloud and/or use both.

Enabling Technologies



Common Features of Cloud Providers



Windows Azure

- Enterprise-level on-demand capacity builder
- Fabric of cycles and storage available on-request for a cost
- You have to use Azure API to work with the infrastructure offered by Microsoft
- Significant features: web role, worker role , blob storage, table and drive-storage

Amazon EC2

- Amazon EC2 is one large complex web service.
- EC2 provided an API for instantiating computing instances with any of the operating systems supported.
- It can facilitate computations through Amazon Machine Images (AMIs) for various other models.
- Signature features: S3, Cloud Management Console, MapReduce Cloud, Amazon Machine Image (AMI)
- Excellent distribution, load balancing, cloud monitoring tools

Google App Engine

- This is more a web interface for a development environment that offers a one stop facility for design, development and deployment Java and Python-based applications in Java, Go and Python.
- Google offers the same reliability, availability and scalability at par with Google's own applications
- Interface is software programming based
- Comprehensive programming platform irrespective of the size (small or large)
- Signature features: templates and appspot, excellent monitoring and management console

Demos

- Amazon AWS: EC2 & S3 (among the many infrastructure services)
  - Linux machine
  - Windows machine
  - A three-tier enterprise application
- Google app Engine
  - Eclipse plug-in for GAE
  - Development and deployment of an application
- Windows Azure
  - Storage: blob store/container
  - MS Visual Studio Azure development and production environment

Cloud Programming Models

The Context: Big-data

- Data mining huge amounts of data collected in a wide range of domains from astronomy to healthcare has become essential for planning and performance.
- We are in a knowledge economy.
  - Data is an important asset to any organization
  - Discovery of knowledge; Enabling discovery; annotation of data
  - Complex computational models
  - No single environment is good enough: need elastic, on-demand capacities
- We are looking at newer
  - Programming models, and
  - Supporting algorithms and data structures.

Google File System

- Internet introduced a new challenge in the form web logs, web crawler's data: large scale "peta scale"
- But observe that this type of data has an uniquely different characteristic than your transactional or the "customer order" data : "write once read many (WORM)" ;
  - Privacy protected healthcare and patient information;
  - Historical financial data;
  - Other historical data
- Google exploited this characteristics in its Google file system (GFS)

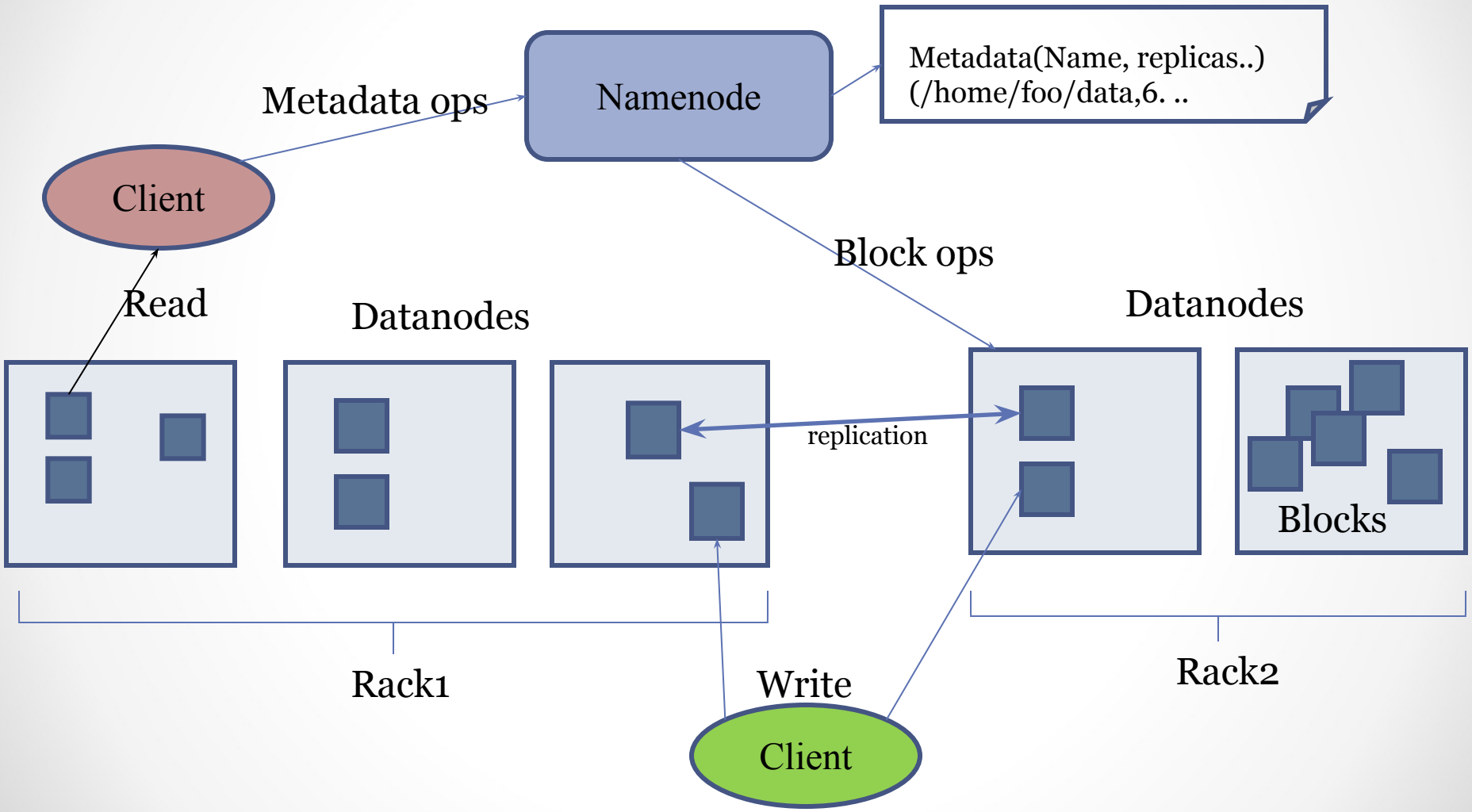
What is Hadoop?

- At Google MapReduce operation are run on a special file system called Google File System (GFS) that is highly optimized for this purpose.
- GFS is not open source.
- Doug Cutting and others at Yahoo! reverse engineered the GFS and called it Hadoop Distributed File System (HDFS).
- The software framework that supports HDFS, MapReduce and other related entities is called the project Hadoop or simply Hadoop.
- This is open source and distributed by Apache.

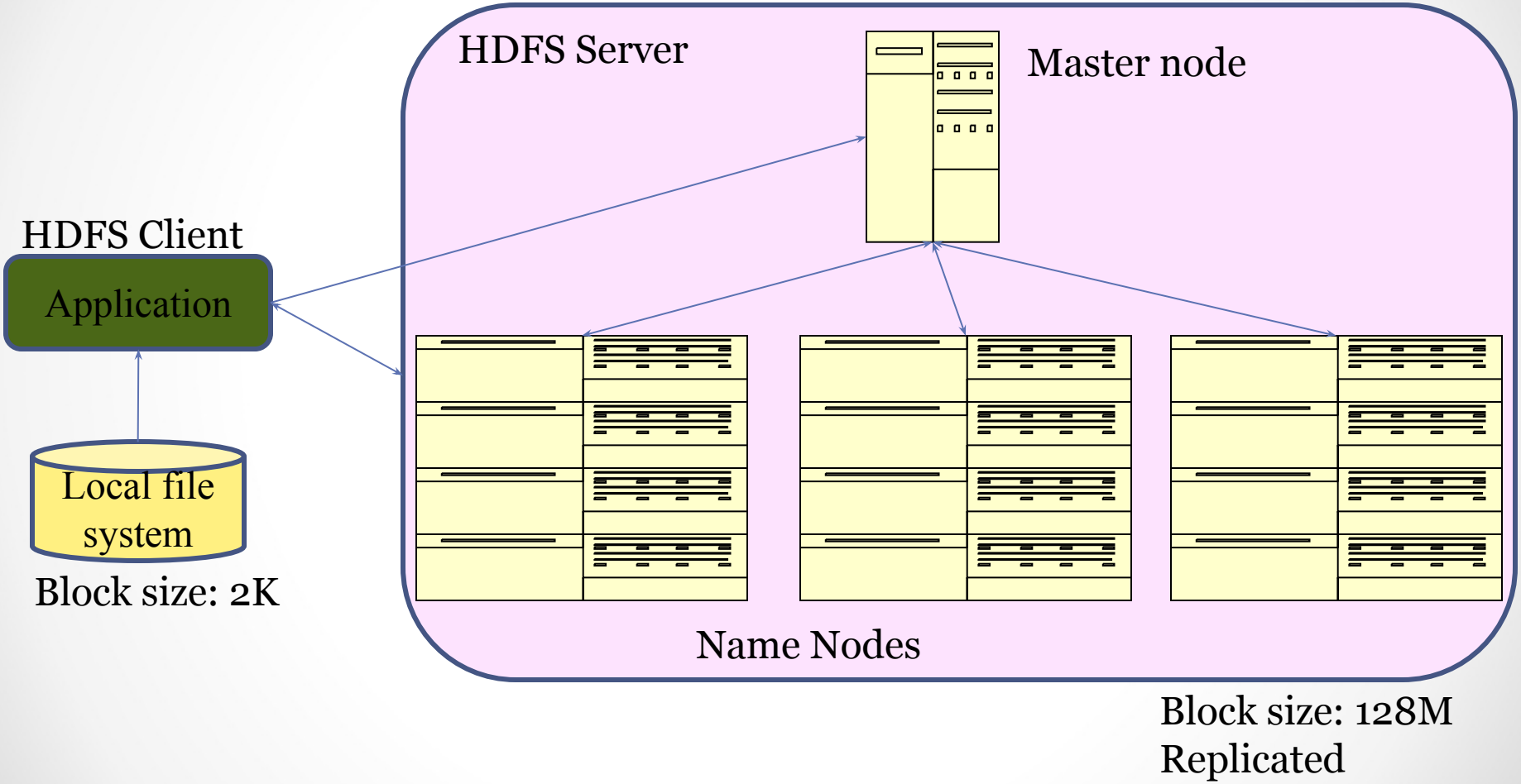
Fault tolerance

- Failure is the norm rather than exception
- A HDFS instance may consist of thousands of server machines, each storing part of the file system's data.
- Since we have huge number of components and that each component has non-trivial probability of failure means that there is always some component that is non-functional.
- Detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS.

HDFS Architecture



Hadoop Distributed File System



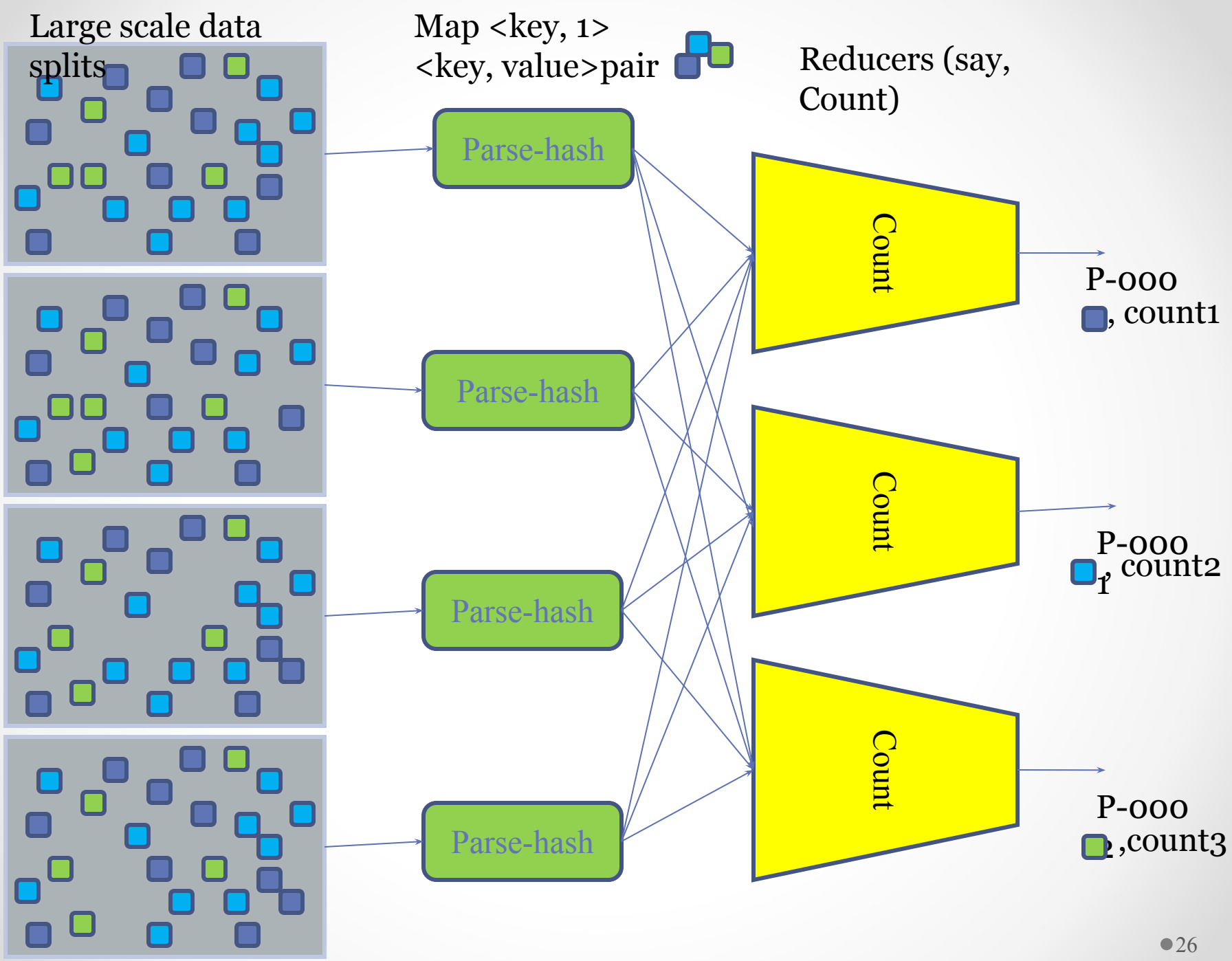
What is MapReduce?

- MapReduce is a programming model Google has used successfully is processing its "big-data" sets (~ 20000 peta bytes per day)
  - A map function extracts some intelligence from raw data.
  - A reduce function aggregates according to some guides the data output by the map.
  - Users specify the computation in terms of a *map* and a *reduce* function,
  - Underlying runtime system automatically parallelizes the computation across large-scale clusters of machines, and
  - Underlying system also handles machine failures, efficient communications, and performance issues.

  -- Reference: Dean, J. and Ghemawat, S. 2008. **MapReduce: simplified data processing on large clusters.** *Communication of ACM* 51, 1 (Jan. 2008), 107-113.

Classes of problems “mapreducable”

- Benchmark for comparing: Jim Gray’s challenge on data-intensive computing. Ex: “Sort”
- Google uses it for wordcount, adwords, pagerank, indexing data.
- Simple algorithms such as grep, text-indexing, reverse indexing
- Bayesian classification: data mining domain
- Facebook uses it for various operations: demographics
- Financial services use it for analytics
- Astronomy: Gaussian analysis for locating extra-terrestrial objects.
- Expected to play a critical role in semantic web and in web 3.0

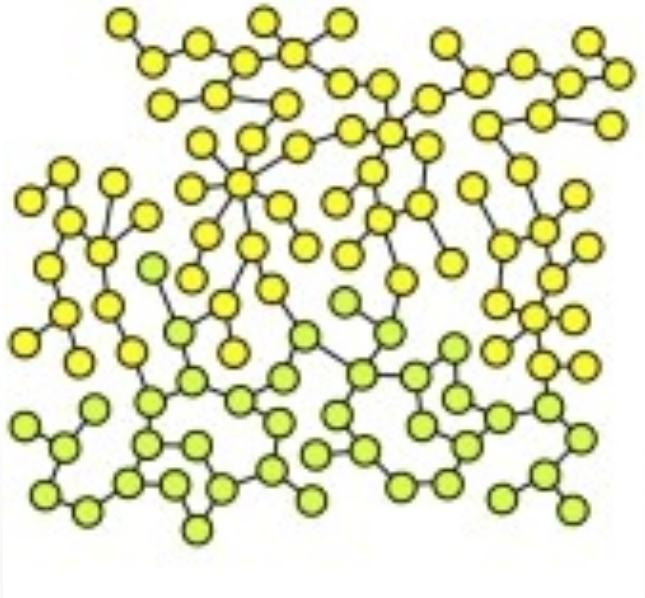

Large scale data splits
Map <key, 1> <key, value>pair
Reducers (say, Count)
Parse-hash
Parse-hash
Parse-hash
Parse-hash
Count
Count
Count
P-000 , count1
P-000 , count2
P-000 ,count3

MapReduce Engine

- MapReduce requires a distributed file system and an engine that can distribute, coordinate, monitor and gather the results.
- Hadoop provides that engine through (the file system we discussed earlier) and the JobTracker + TaskTracker system.
- JobTracker is simply a scheduler.
- TaskTracker is assigned a Map or Reduce (or other operations); Map or Reduce run on node and so is the TaskTracker; each task is run on its own JVM on a node.

Demos

- Word count application: a simple foundation for text-mining; with a small text corpus of inaugural speeches by US presidents
- Graph analytics is the core of analytics involving linked structures (about 110 nodes): shortest path



A Case-study in Business: Cloud Strategies

Predictive Quality Project Overview

Problem / Motivation:

- Identify special causes that relate to bad outcomes for the quality-related parameters of the products and visually inspected defects
- Complex upstream process conditions and dependencies making the problem difficult to solve using traditional statistical / analytical methods
- Determine the optimal process settings that can increase the yield and reduce defects through predictive quality assurance
- Potential savings huge as the cost of rework and rejects are very high

Solution:

- Use ontology to model the complex manufacturing processes and utilize semantic technologies to provide key insights into how outcomes and causes are related
- Develop a rich internet application that allows the user to evaluate process outcomes and conditions at a high level and drill down to specific areas of interest to address performance issues

Why Cloud Computing for this Project

- Well-suited for incubation of new technologies
  - Semantic technologies still evolving
  - Use of Prototyping and Extreme Programming
  - Server and Storage requirements not completely known
- Technologies used (TopBraid, Tomcat) not part of emerging or core technologies supported by corporate IT
- Scalability on demand
- Development and implementation on a private cloud

Public Cloud vs. Private Cloud

Rationale for Private Cloud:

- Security and privacy of business data was a big concern
- Potential for vendor lock-in
- SLA's required for real-time performance and reliability
- Cost savings of the shared model achieved because of the multiple projects involving semantic technologies that the company is actively developing

Cloud Computing for the Enterprise

What should IT Do

- Revise cost model to utility-based computing: CPU/hour, GB/day etc.
- Include hidden costs for management, training
- Different cloud models for different applications - evaluate
- Use for prototyping applications and learn
- Link it to current strategic plans for Services-Oriented Architecture, Disaster Recovery, etc.

References & useful links

- Amazon AWS: http://aws.amazon.com/free/
- AWS Cost Calculator: http://calculator.s3.amazonaws.com/calc5.html
- Windows Azure: http://www.azurepilot.com/
- Google App Engine (GAE): http://code.google.com/appengine/docs/whatisgoogleappengine.html
- Graph Analytics: http://www.umiacs.umd.edu/~jimmylin/Cloud9/docs/content/Lin_Schatz_MLG2010.pdf
- For miscellaneous information: http://www.cse.buffalo.edu/~bina

Summary

- We illustrated cloud concepts and demonstrated the cloud capabilities through simple applications
- We discussed the features of the Hadoop File System, and mapreduce to handle big-data sets.
- We also explored some real business issues in adoption of cloud.
- Cloud is indeed an impactful technology that is sure to transform computing in business.