Big Data and the Cloud
Trends, Applications, and Training

Stavros Christodoulakis
MUSIC/TUC Lab
School of Electronic and Computer Engineering
Technical University of Crete
Data Explosion

Data becomes available at a rapid pace

• Forrester estimates that data volume doubles every 18 months for mission critical applications

• Difference: real time data from business processes for decision making

• Web site tracking, customer goals, market segments, CRM,..

• Social networking comments, postings, tweets, pictures,..
Data Explosion

• Mobile applications, location based and personal data

• **Sensor data, Internet of Things, Web of Things**

• One estimate says that 30 Billion devices will be connected in the internet of things by 2020

• DNA sequencing, Space data,...
Big Data Acquisition

Tracking of

– business processes

– products, shipments, distributions

– behavior, activities and distribution of customers

– public assets (road network, water resources, etc.)

– environment, atmosphere, weather, pollution, ...
Big Data Use

Explosion of interest due to
- dramatic drops in the cost of hardware and the increase of capacity and speed
- Proliferation of new sensing devices
- Recognizing their value for real time decision making

Decision makers should base their real time decisions on data not intuition only
- Evaluate, improve react in real time in their business process

Making sense of diverse data, understanding customer behavior, define customer segments, offer competitive services..
Big Data Analytics

Analytics involves

• **building** models

• **training** the models and estimating their parameters

• **validating** the models

• **applying** them in the problem domain

• **visualizing** the results
Big Data Analytics

• Categories of Analytics
  – **Descriptive**: model the past behavior
  – **Predictive**: forecast based on available data
  – **Prescriptive**: assess actions, assist decision making

• Analytics tools may involve:
  – Data mining, text mining
  – Statistical and quantitative analytics
  – Predictive analytics tools
  – Data Visualization tools
Big Data Challenges

• **Variety**: data types, data integration

• **Velocity**: data production, change, continuous data tracking, speed of interaction

• **Volume**: archival speed of access

• **Veracity**: data reliability and trust

• **Value**: data exploitation for profit,…
Cloud Computing

Cloud computing delivers computing services, data storage, computation, networking, to users through internet infrastructure and standards (service oriented computing)

• Services are offered at any location, any time

• Scalable services (Big Data)

• Services of any quantity that the users want

• Costs based on the resources used
Advantages of Cloud Computing

• Drives computing resources to commoditization and price competition

• Resources are always available

• Services paid according to use. Offers availability and reliability of computing through Service Level Agreements

• Global user reach: services are accessible by web mobiles, etc., at any point, any time
Cloud Offerings

• **IaaS**: Infrastructure as a Service: Offers services of computing resources

• **PaaS**: Platform as a Service: Offers Services of development tools

• **SaaS**: Software as a Service: Offers Services of Software applications
Infrastructure as a Service

• Large amount of computing resources to satisfy requests for services from the internet

• **Virtualization services** allow to pull together physical resources to satisfy the needs of service requests

• **Server virtualization** functionality abstracts the physical resources and presents them as virtual machines that appear to the application and users as a physical system – Servers may run different OS’s

• **Hypervisors** are management layers that facilitate launching of virtual machines from the virtual disk (Hyper-V,..)
Infrastructure as a Service

• **Storage virtualization** distributes redundant files or blocks across physical storage
  – Load balancing, pricing billing facilitated
  – Many companies offer storage services for robustness, scalability, reliability, availability, replication control

• **Amazon Elastic Computing (EC2)** is IaaS accessed through REST and SOAP service interfaces.
  – Provides Elastic Block Storage with replication

• Major IaaS providers: Amazon, Google, Microsoft
Storage and Processing Services

• **Tiered Storage Architectures** for archival nature applications in the cloud, big data types (broadcasting, video, space data), back up,..
  – 1. PCI Flash Storage, transfer rate 1500MB/sec at 26 US$ per GB
  – 2. SSD Solid State Drives, t.r. 500MB/sec at 2 US$ per GB
  – 3. SAS SCSI Disks t.r. 200 MB/sec at .70 US$ per GB
  – 4. SATA Disks and Tape Drives, t.r. 140-150 MB/sec at .04 US$ per GB

• Long archival life in new optical disks (preservation, film industry)

• Data placement for system and application data important problem
Storage and Processing Services

Hadoop has become a dominant open source Framework for storage and large scale distributed processing of large data sets in commodity hardware.

HDFS: distributed filing system, services blocks

YARN: resource manager, tries to place processing tasks near the data, cluster coordinator allocates tasks
MapReduce has become an important programming model for processing Big Data and Cloud applications

- It gives a programmatic model for distributing and parallelizing heavy data processing jobs
- Input is split in appropriate sizes, described with key/value pairs and distributed for parallel map processing.
- Output is in different key/value pairs and sent to a different set of processors for summarizing
- Redundant copies are sent, scheduler checks the progress, for reliability
Storage and Processing Services

• Implementation of Map/Reduce in most platforms

• Elastic MapReduce is an implementation on the Amazon EC2 Storage Cloud Services

• Tradeoff of communication and processing costs

• Criticism for lack of innovation, no complex query processing, learning a new language
Storage and Processing Services

- Amazon Elastic MapReduce includes support for large data bases stored on Hadoop file system with SQL-like language and full Map/Reduce
  - No transactions, limited subquery
- Hive: SQL-like offering on top of Hadoop using MapReduce
- Impala: SQL-like on Hadoop on Share Nothing
- HAWO: dbms optimization, HDFS to give work to dbms workers
Data Integration

Data Integration is often a major issue for Big Data Analytics

- Information integration from diverse data types and languages

- The eXtreme Analytics Platform (XAP) supports analytics processing from multiple structured and unstructured sources
  - Runs on a modified version of Hadoop, uses a script language that is converted to MapReduce

- Business Process Execution Language (BPEL), a SOAP Services standard, has been proposed as a language for coordinating data exchange in clouds, passing references to data between services to guarantee correct processing
Continuous Analytics Support

- Applications: weather predictions, stock quotes, steams,..

- Stream processing Frameworks can be deployed on Cloud offerings

- Continuous Analytics as a Service extends DBMS models to provide continuous services through an SQL-like interface to static and streams of data

- SAP HANA One provides real-time analytics for SAP applications on AWS
  - In memory platform, monthly subscription
Continuous Analytics Support

• **Storm**: real time stream processing Framework based on data flow programming
  – In contrast to Map/Reduce which is batch processing

• Storm applications are designed as a DAG where edges are streams and direct data from a node to another

• Processes run indefinitely, until they are killed

• storm-deploy aims to make Storm available on AWS EC2
Database Trends

• The database market is healthy

• 30 Billion US$, projected 35 Billion by 2017 (Forrester)

• OLTP and DW grow 10% per year
Relational OLTP

- Relational OLTP target to improve performance using Scale Out Architectures

- Scale out Architectures are share nothing architectures using many servers

- Use horizontal partitioning of tables to place data from different tables on the same server (sharding)

- Cheaper by far in comparison to scale up, better reliability

- More complex application software for the sharding, and security issues

- Recent advances aim to automate sharding, resharding, load balancing
Enterprise Data Warehousing (EDW)

• EDWs store data for business intelligence, analytics, etc.

• Use Extract Transform Load (ETL) to move from OLTP to DW

• DW vendors move to offer appliances: purpose built applications tuned to specific environments and workloads.
  – One button deployment, simplified maintenance, support, virtualization, availability, high interconnect,…

• EDWs move to NoSQL, Graph Databases, KeyValue Stores, Document Stores
Graph Databases

• Speedup access to data having many relationships

• Applications in social networks, Facebook, Twitter, LinkedIn, recommendation engines, dependency analysis, etc.

• Neo4j, AllegroGraph, IBM DB2 NoSQL, Graph Store,...
Key Value Stores

• Store key and value pairs

• Can store dynamic number of key value pairs per record

• Fast access to distributed data

• Leave out some SQL features

• DynamoDB (simple key value), Apache Cassandra, Amazon, IBM, Oracle
Document Stores

• Schemaless, records have variable types and many attributes

• Columns can have more than a value

• Nested structure of records

• Apache Couch DB, MarkLogic Server, MongoDB,..
Object Databases

- Tuned to object programming environments

- Objectivity, GemStone,..
Specialized Databases

• Include mobile, cloud, in memory, standalone

• Cloud data bases automate the provisioning, administration, backup, recovery, availability, security, scalability

• No need for data base administrator, backups,..

• Economies of scale through elastic computing
Specialized Databases

- Database as a Service (AWS RDS, simplified DBMS)

- Amazon offers Oracle and Microsoft SQL Server as managed virtual machines

- Amazon Dynamo DB (key value)

- Oracle, Microsoft, Salesforce offerings in the cloud
Network Services and the Cloud

• The Cloud uses internet for offering the services of resources

• This adds orders of magnitude of delay for accessing the data than accessing it through local area nets

• Microsecond access VS tens of milliseconds for going across the US (30 ms only due to speed of light)
Network Services and the Cloud

• Within organizations:
  – Bandwidth of switches 50 Gb/s
  – Personal capacity 1Gb/s

• Border routers for enterprises: 1-10 Mb/s speed depending on the length of cables

• Moving services to cloud represents bandwidth reduction of about 1000 times

• Amazon cost calculations show that even with 10-100 Mb/s bandwidth costs are 75% to 99% of the average bill

• Some predictions that bandwidth supply and demand differences will grow
Distributed Clouds

• Distributed Clouds is the only way to reduce bandwidth demand

• Move data and computation near to consumption

• Programmable private networks can dynamically adjust the data flow over the physical network

• The OpenFlow Protocol (Stanford) permits applications to reprogram the network during the course of the application

• Allow the network to recognize the application and give the agreed services
Federated Distributed Clouds

GENI network (NSF)

• 50 Clouds at Universities and R&D Centers in the US

• A Cloud has 80-100 cores and terabytes of storage

• Cloud is programmable. Can allocate virtual machines anywhere in the net, specify precisely how they interconnect, traffic priorities, etc.
Platform as a Service (PaaS)

**PaaS** acts as a run time environment that supports a development and on-line collaboration

- Development environments, integration services, workflow facilities, HTML, JavaScript, visualization tools, collaboration services,..

- Services for developers (as opposed to administrators of IaaS)

- Google Application Engine, Microsoft Azure, Salesforce,..
Software as a Service (SaaS)

SaaS provides services directly consumable by end users (as opposed to developers)

- Services like ERP, CRM, etc., are centrally managed and updated

- A problem is that they offer a complete functionality based on a model (of the vendor) that may not be the business model of the customer

- Salesforce, NetSuite, healthcare solutions, transport, logistics, etc.
Cloud Native Workloads

- Data serving, search, social, mobile apps, batch processing
- Web apps (web 2.0), rich internet apps (videos, games)
- NoSQL and HPC for scientific apps
- Batch processing like data mining, BI, disaster recovery, development, testing
- Elasticity and transient usage requirements
Trends and Directions

• Current Public Cloud offerings highly emphasize the fast development of virtual machines as inexpensively as possible

• They are not strong in providing service automation, orchestration, management of workflows

• IaaS is strongly dominated by Amazon, Google, Microsoft, huge investments, difficult to compete

• Emphasis should be placed on higher service layers and their tight integration
Trends and Directions

• Today clouds are highly centralized, non sustainable.

• Develop NaaS giving virtualization to Networks

• Develop distributed federated Clouds moving the execution near the data

• Offer Database as a Service, Messaging as a Service, Identity as a Service, Network as a Service

• Database tuning, self healing, automatic notifications, workflow management, ETL, Hadoop as a Service

• IT as a Service, Service Ontologies
Trends and Directions

• Visualization in the cloud is problematic because the cloud acts as a batch model of computation and network connections are too slow for interaction
  – Use sampling
  – Make map reduce, and cloud dbms’s interactive
  – Iterative facetted explorations in the cloud

• Dashboard adaptation, domain visualization
Graduate Level Research

• TUC has several researchers in Big Data and Cloud related topics

• **ECE School information systems**
  – Data bases, map/reduce, sensor networks, Storm, security, heterogeneous data processing, federated architectures, medical, biomedical, biodiversity applications,..

• **Environmental Engineering School**
  – Water resources, pollution, weather prediction, traffic, energy,..

• **Industrial Engineering (business data)**

• **Mineral Resources (3D earth data, seismic data, space data)**
Graduate Level Research

• Difficult in Universities that do not have rich cloud infrastructure and apps

• New Cycles of R&D [J. Zysman, UC Berkeley, describing the approach of the top American Technical Universities]:
  – In a fast moving cloud era there can not be a clean separation between research strategies and innovation strategies.
  – Cloud developments occur in collaboration. Requires fast product delivery, testing, maintenance
  – Filling in the research projects as they go along
  – EU Marie Curie Training programs with industry
MEng Training

- TUC has research oriented MEng, PhD programs

- What is the training that should be given to engineers in 1.5 to 2 years of residence or distance learning to meet the needs of the industry? Distance learning with our platform?

- Depends on the country..

- Looked at major sites in the US offering jobs and the skills that require

- Major by far demand for software engineering, especially in the areas of Web Applications, Web Services, for years.

- Analytics demand started moving fast, but they may need also business knowledge, data management and software
MEng Training
MEng Training

Big Data, Web and Mobile Services

• Course areas:
  – Big data and Analytics
  – Web Application Development (SE)
  – Service Oriented Engineering
  – Cloud Technologies, Algorithms and Architectures
  – User Interfaces and Visualization
  – Mobile Computing
  – Security

• Selections mostly from other Schools
• Selections in advanced probability and statistics
• Collaborative project with integrated development (4 months)
Conclusions

• Big Data and Cloud Computing has significant advantages for the industry

• There is a long way to go for offering integrated and Big Data and Cloud Services

• The industry needs should influence research and training in the Universities