New Trends in Database Systems

Ahmed Eldawy
Project Timeline

10/13
Initial project idea

11/8
Report outline is due

12/1
Final deliverables are due

10/25-10/27
Proposal presentation

11/17-12/1
Full presentation

12/5 [8-11 am]
Final presentation
Project Grading

- Thursday 10/13: Initial project idea
  - Agree on goals and deliverables
- Grading will be based on our agreement

- Proposal presentation 15%
- Report outline 5%
- Full presentation 15%
- Deliverables 15%
- Final presentation 10%
- Project total 60%

9/27/2016
Presentations

- Two presentations per class
- Review slides in office hours
- Email your slides before class
Big Data Management
Big Data

Straight Ahead
All of the information

Information you need!
Big Data, Big Impact: New Possibilities for International Development

The amount of data in the world is exploding - a large portion of this comes from the interactions over mobile devices being used by people in the developing world - people whose needs and habits have been poorly understood until now. Researchers and policymakers are beginning to realize the potential for channeling these torrents of data into actionable information that can be used to identify needs & provide services for the benefit of low-income populations. This discussion note is a Call-to-action for stakeholders for concerted action to ensure that this data helps the individuals and communities who create it.
Interest in Big Data in the US

■ March 2012: Obama administration unveils BIG DATA initiative: $200 Million in R&D investment

FOR IMMEDIATE RELEASE
March 29, 2012

OBAMA ADMINISTRATION UNVEILS “BIG DATA” INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a “Big Data Research and Development Initiative.” By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some of the Nation’s most pressing challenges.

■ June 2013: Washington Post is calling Obama “The Big Data President”
Interest in Big Data in Europe

March 2014: David Cameron and Angela Merkel talking about Big Data in a Computer Expo in Hannover, Germany
Four V’s of Big Data

**Volume**
- Scale of Data
- It’s estimated that 2.5 quintillion bytes (1.33 quadrillion gigabytes) of data are created each day.
- Most companies in the U.S. have at least 100 terabytes (100 billion gigabytes) of data stored.

**Velocity**
- Analysis of Streaming Data
- By 2016, it is projected there will be 18.9 billion network connections—almost 2.5 connections per person on earth.

**Variety**
- Different Forms of Data
- As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (161 billion gigabytes).
- By 2014, it’s anticipated there will be 420 million wearable, wireless health monitors.
- 4 billion+ hours of video are watched on YouTube each month.
- 400 million tweets are sent per day by about 200 million monthly active users.

**Veracity**
- Uncertainty of Data
- By 2015, 4.4 million IT jobs will be created globally to support big data, with 1.9 million in the United States.

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC2, GDC, IBM, HIPAA, Q4Q

9/27/2016
Big Data Vs Big Computation

- Full scans (e.g., log processing)
- Range scans
- Point lookups
- Iterations
- Joins (self, binary, or multiway)
- Proximity queries
- Closures and graph traversals
Market Size

- Applied Market Research: **Hadoop Market** (SW, HW & Services) - $2.0B in 2013 to $50.2B by 2020
- March 2014 evaluation of **Cloudera**: $4.2B
- **IDC**: Big data storage revenue > $6B in 2014
- **Merv Adrian**: “Gartner’s quarterly Hadoop webinar in Feb 2015 showed that adoption of Hadoop is not rising quite as dramatically as some might believe. It’s flat compared to Q42014.”
Big Data Applications

› Web search
› Marketing and advertising
› Data cleaning
› Knowledge base
› Information retrieval
› Internet of Things (IoT)
› Visualization
› Behavioral studies
Publicly Available Datasets

- Data.gov
- Data.gov.uk
- Twitter Streaming API
- Yahoo! Webscope [http://webscope.sandbox.yahoo.com/]
- GDELT [http://www.gdeltproject.org/]
- Instagram API
Big Data Landscape 2012

Big Data Landscape

Infrastructure
- NoSQL / NewSQL Databases
  - 10gen
  - VoltDB
- Hadoop Related
  - cloudera
  - HADAPT
- Crowdsourcing
  - CROWD COMPUTING
- Storage
  - EMC
- Management / Monitoring
  - Dell

Analytics
- Analytics Solutions
  - OPERA
  - Palantir
- Data Visualization
  - Quid
  - Pivot
- Statistical Computing
  - SKYTERE

Applications
- Ad Optimization
  - m6d
  - DataXu
- Marketing
  - Turn

Data Sources
- Data Marketplaces
  - factuaL
- Data Sources
  - knoema

Open Source Projects
- Hadoop
- Cassandra
- Apache
- Storm

© Matt Turck (@mattturck) and ShivonZilis (@shivonz)

9/27/2016

Agenda

- Motivation
- Definition
- Applications
- Components
- Systems
Storage of Big Data

- Data is growing faster than Moore’s Law
- Too much data to fit on a single machine
- Partitioning
- Replication
- Fault-tolerance
Hadoop Distributed File System (HDFS)

- The most widely used distributed file system
- Fixed-sized partitioning
- 3-way replication
- Write-once read-many
Indexing

- Data-aware organization
- Global Index partitions the records into blocks
- Local Indexes organize the records in a partition
- Challenges:
  - Big volume
  - HDFS limitation
  - New programming paradigms
  - Ad-hoc indexes
Fault Tolerance

- Replication

- Redundancy

- Multiple masters
Streaming

- Sub-second latency for queries
- One scan over the data
- (Partial) preprocessing
- Continuous queries
- Eviction strategies
- In-memory indexes

9/27/2016
Task Execution

- MapReduce
  - Map-Shuffle-Reduce
  - Resiliency through materialization
- Resilient Distributed Datasets (RDD)
  - Directed-Acyclic-Graph (DAG)
  - In-memory processing
  - Resiliency through lineages
- Hyracks
- Stragglers
- Load balance
Query Optimization

- Finding the most efficient query plan
- e.g., grouped aggregation

Cost model (CPU – Disk – Network)
Provenance

- Debugging in distributed systems is painful

- We need to keep track of transformations on each record
Big Graphs

- Motivated by social networks
- Billions of nodes and trillions of edges
- Tens of thousands of insertions per second
- Complex queries with graph traversals
Agenda

› Motivation
› Definition
› Applications
› Components
› Systems
Hadoop Ecosystem

- Pig
- Hive
- Giraph
- Mahout

MapReduce Query Engine
Yet Another Resource Negotiator (YARN)
Hadoop Distributed File System (HDFS)
Spark Ecosystem

- Spark SQL
- Data Frames
- MLlib
- GraphX
- SparkR
- Spark Streaming

Resilient Distributed Dataset (RDD) a.k.a Spark Core

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS)
AsterixQL

AsterixDB

HiveQL

HiveSterix

PigLatin

Other compilers

MapReduce Jobs

Hadoop MapReduce Compatibility

Pregel Jobs

Pregelix

Hyracks jobs

Hyracks Data-parallel Platform

Algebricks Algebra Layer

Other compilers

MapReduce Jobs

Hadoop MapReduce Compatibility

Pregel Jobs

Pregelix

Hyracks jobs

Hyracks Data-parallel Platform

AsteixDB

HiveSterix

Other compilers

MapReduce Jobs

Hadoop MapReduce Compatibility

Pregel Jobs

Pregelix

Hyracks jobs

Hyracks Data-parallel Platform
SpatialHadoop

Applications: SHAHED – MNTG – TAREEG

- Language: Pigeon
- Visualization: HadoopViz
- Operations: Basic operations – CG_Hadoop
- Indexing: Grid – R-tree – R+-tree – Quad tree
Conclusion

- Big data is a hot research area
- Aspects of big data (V’s)
- Database research components in big data
- Available open source big data systems
- A good project ideas
  - Build an application using one of the big data systems
  - Experimentally evaluate different big data systems
Thank You!