BREAK OUT SESSION

Big Data Systems

Session Leaders
Magdalena Balazinska & Kunle Olukotun
Overarching Themes in this Area

• Democratizing big data
• Data acquisition and cleaning
• Making complex analytics fast
• Data velocity
  – Need to process data streams from IoT, video, other
• Data variety
  – Need to process graphs, structured, unstructured, multimedia
• Data volume
  – Must integrate and analyze immobile data, distributed around world
• Reproducibility, long-term preservation, and sharing
Recent Successes (last 3 years)

• Increasingly efficient, open-source systems
  – Spark, Impala, Myria, Asterix, GraphLab, etc.
• Growing cloud service offerings
  – Data management and also ML services
• Growing availability of ML algos and datasets
• Knowledge bases
• Systems that go from high-level DSLs to hardware-specialized implementations
• Big science projects (e.g., LHC and SDSS/LSST)
• Tools for data science and collaboration
Major Obstacles Impeding More Rapid Progress

- Data science education across domains
- Cloud services can be hard to use cost-effectively
- ML and DB remain poorly integrated
- We settled on commodity but need to explore other architectures
- Need to unify abstractions
  - Big data is a mix of relational algebra, linear algebra, ML, etc.
- Data science is a high-touch business
  - How to choose ML algo? Tune data analysis pipelines?
  - Can we have even higher-level interfaces for data science?
  - Data in many different formats
- Data correctness, corruption, long-term preservation
- Hard to share:
  - Create metadata automatically
  - Make data not only available but easily accessible
  - Risks associated with data sharing (burden, responsibility, scooped)
Areas that Need More Attention

• Cross-disciplinary data science education
  – Across levels undergraduate, graduate, master’s
• Storage remains the bottleneck
• Compute
  – Future of hardware is increasingly heterogeneous
  – but still no abstractions for shielding complexity
• Cross-stack innovations:
  – PL, compiler, database, OS, networking, hardware
• End-to-end analysis pipelines
  – Need to support users end-to-end
• Reproducibility, sharing, and reuse
• Long-term curation and preservation
Strategic Priorities & Investments That Will Advance Innovation

• *Democratizing Big Data*
  
  – Productivity tools and methods
    • End-to-end data science pipelines
    • Easier-to-use cloud analytics systems
    • Cost-effective cloud analytics
  
  – Expressing complex analysis
    • Data management + ML + ...
    • Also leverage legacy code
    • Common analytic frameworks (laptop to cluster/clouds)
    • Higher-level interfaces to data analytics
      – SQL, visualizations, natural language, other?
  
  – Correctness and auditability
  
  – Applications of data science
Strategic Priorities & Investments That Will Advance Innovation

• **Reproducibility**
  – Data sharing and preservation
  – Code sharing and preservation
  – Responsibility and ethics of data analysis
  – Long-term preservation

• **Infrastructure**
  – A data observatory (a single, logical place)
  – Partner with cloud providers
  – Leverage existing HPC centers
  – Explore what is the best, global approach
Strategic Priorities & Investments That Will Advance Innovation

• *Data acquisition and cleaning*
  
  – Data cleaning and integration
  – Managing probabilities, errors, approximations
    • Data is not always precise: density distributions
    • Computation/analysis uses approximations
Strategic Priorities & Investments That Will Advance Innovation

• *Making complex analytics fast*
  – Interactive analytics
  – Innovation in architectures
  – Across-the-stack innovations
  – Benchmarks: data sets, analytics, etc.
  – In-memory analytics
  – Complex analytics
  – Mobile devices or even IoT devices
  – Federated analytics
  – HPC + dataflow systems
Strategic Priorities & Investments That Will Advance Innovation

• **Data velocity**
  – Stream processing

• **Data variety**
  – Different types of data structured, unstructured, etc.

• **Data volume**
  – Manage data value over time
  – Analysis over data distributed across data centers