Outline:

* 1.2 Visualizing Twitter
  + GPU Accelerated Database similar to MapD
    - Visualizations can be served to users
    - Interactivity easier to attain
    - Filtering in parallel
    - Analysis can be done in SQL but plumbing in parallel
* 2.1 Data Cleaning Integration
  + Data Tamer – Schema Integration
    - Use APIs for normal data sources, e.g. Slack & email servers
    - Map local to global structures for disparate datasets
* 2.2 Data Cleaning Integration 3 Years Later
  + Data Civilizer? Unsure
* 2.3 Hosted Data Platforms & the Cloud
  + Train ML models on Slack data
    - Embeddings from all users; specific embeddings per person on top
  + Storage
  + Serving ML models
  + Discuss motivation: variable utilization, economies of scale, cost associativity
  + When to use SaaS, PaaS, and IaaS
  + Whether private or public cloud
  + Access interface (e.g. Open or Proprietary)
  + Security on cloud
    - Control security properties (encryption, key rotation, access rolees)
    - Provider compliance (e.g. HIPAA)
    - Cryptographic advances
* 3.1 Modern Databases
  + Column storage databases + benefits of using (50-100X faster, compression)
  + Main memory DB for OLTP
  + Array DBMSs for complex analytics
    - One liner queries and model training, e.g. clustering
    - E.g. SciDB – use of built-in functions
    - Graph DBMS for R&D
  + Hadoop for simple historical search of conversations (embarassingly parallel)
    - Hive – Hadoop – HDFS – nodes
    - Use as R&D as well
    - Only get one shot, use for right case
  + Hire a good Chief Data Officer
* 3.2 Modern Databases 3 Years Later
  + Volume (data science/BI), Velocity, Variety (diversity/integration)
  + Which C-store to use (e.g. not Oracle), which products are good but hard to find?
  + Data science will replace BI (timeline, transition period, different folks)
  + Array DBMS again
  + Supporting data science 1
    - Main memory (Spark)
    - Move query to the data
    - UDFs to support fast and efficient analytics (Spark)
    - R&D to support UDF dev
  + Supporting data science 2
    - Array DBMS
    - In- DB analytics
    - No table-to-array conversion (SciDB)
  + Try both til a winner is in place, small scale POCs
* 3.3 Distributed Computing Platforms
  + Applications (ETL, action recommendation, ad-hoc queries, log mining)
  + Preprocess text data for ML via distributed computing
  + Convergence with SQL – Async computing – Stream processing
* 4.1 Security
  + Security as a negative goal – limit attack surface
  + Encryption
    - Encrypted queries, CryptDB, tradeoff some generality for security
    - Fully homomorphic, mOPE
* 4.2 Multicore Scalability
  + Sharing vs. partitioning
* 4.3 Visualization & User Interfaces
  + “The. Use of computer-generated, interactive, visual representations of data to amplify cognition.” – Mackinlay & Schneiderman 1999
  + Pre-attentive viz
  + Nominal, ordinal, quantitative, relational data types (see page 14)
  + Allowance for interactivity - Direct manipulation – Exploration
  + Tufte graphical integrity (effect of graph / effect in data)
  + Exploratory vs. Confirmatory examples
  + Advanced users vs. Amateurs (p. 33) – tradeoff of direct manipulation
  + Interaction strategiese (pan/zoom, filter, facet, aggregated views, templates)
    - Critique of hierarchical browsing
  + Data backed website