Design Principles for Effective Knowledge Discovery from Big Data

WICSA 2012, August 22-nd, 2012

Edmon Begoli, James Horey – Oak Ridge National Laboratory

Outline

- Introduction
- Background and Motivation
- Principles
- Implementation Notes and Results
- Future work

Why this topic matters?

- Unprecedented growth of data being produced, collected and analyzed
- Big data movement – velocity, variety and volume
- Field of data management is in flux*
- Patterns, data access standards and best practices are yet to emerge

*See "NewSQL" debate http://goo.gl/VqcXA and related http://goo.gl/UfEKH

Introduction

- What is Knowledge Discovery (KD)?

  Discovering new knowledge from existing data sets by applying computational and cognitive methods for analysis and comprehension

- Importance of Knowledge Discovery and emerging challenges

  Making sense of of large volumes of heterogeneous data is serendipitous, challenging process that is best facilitated through well defined, disciplined and theoretically founded process

- KD Process encompasses

  domain understanding, data collection, data organization, data analysis, review of findings and iterative revisions

Big picture …

OAK RIDGE National Laboratory

Support heterogeneity, comprehensiveness and flexibility of data analysis

Design Principles

- Principle 1: Support a Variety of Analytic Methods
- Principle 2: One Size Does Not Fit All
- Principle 3: Make Data Accessible

Principle 1: Support a Variety of Analysis Methods - Comprehensiveness

- Knowledge discovery from "Big data" requires iterative application of various analytical methods performed by several different teams of different specialties – statisticians, machine learners, data miners, mathematicians, domain experts
- Analysis will often be **exploratory** (in addition to traditional confirmatory) hence architecture should support
  - Statistical Analysis
  - Data Mining and Machine Learning
  - Visualization
  - Geospatial Analysis
- Make available **diverse** and appropriate tools and organize data accordingly

Principle 2: One Size Does Not Fit All* - Heterogeneity

- Need for heterogeneous architecture that supports **different types of data** (structured, semi-structured and unstructured), **different data structures** as well as **different kinds of analysis** (see principle 1) **and modes of analysis** (batch, streaming, interactive)
- Leverage Hadoop-like and other hybrid platforms or build your own that offers linearly scalable "swiss knife" capabilities
- Don't limit yourself and your choices by the architecture

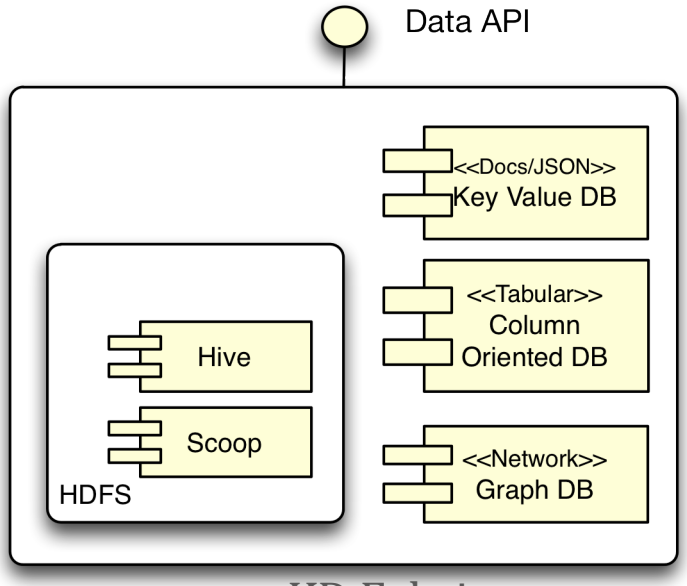
*title inspired by: "**One size fits all**: an idea whose time has come and gone", M. **Stonebraker** - Communications of the ACM, 2008

Principle 3: Make Data Accessible - Accessibility

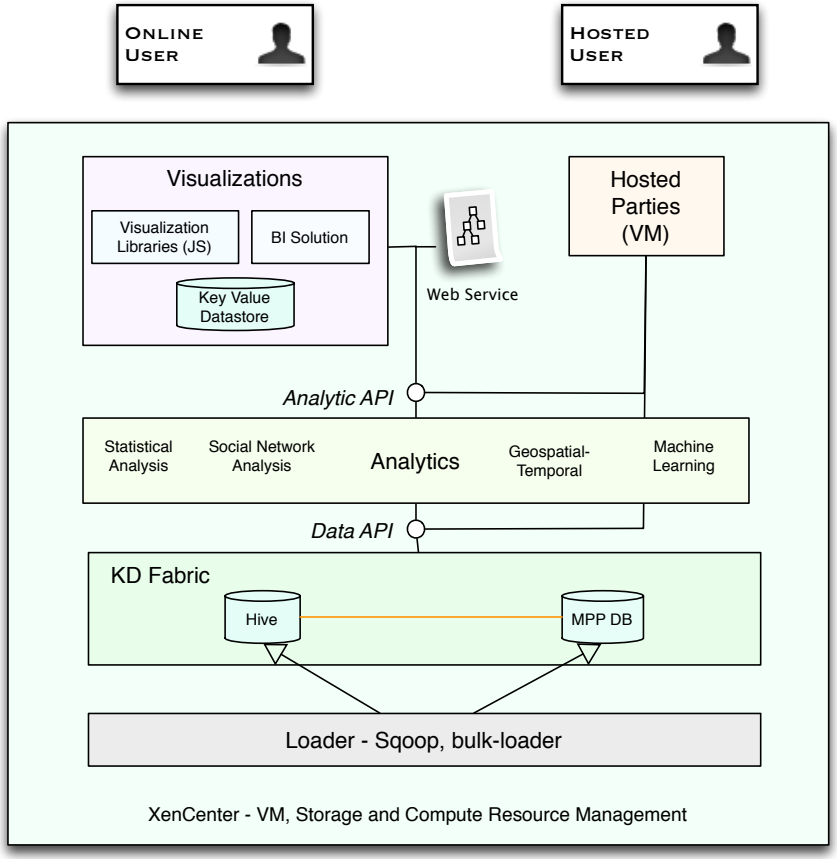
- Make data accessible (liberalize it) by:
  - 1) Use of open, popular standards - accessibility
  - 2) Use of lightweight architectures - scalability
  - 3) Expose results using an API - uniformity
- Open data to as many consumers as possible by using popular standards for data access and representation (e.g. REST, JSON, OData,…)

Implementation Case Study – ORNL Knowledge Discovery Infrastructure (KDI)

- KD Fabric and Analytic Services
- Data and Analytic APIs
- Visualization



KD Fabric



KDI

Results

- Analysis performed in Java/MapReduce, Python, SQL/ HQL and R on the same platform by the same team
- For the first time ever, detailed analysis performed and discoveries occurred over two major national healthcare datasets
- Agility of analysis
  - installation, comprehension, analysis – 4 weeks to major new discoveries
  - recovery of knowledge discovery operations – 2 days
- Successful application of major machine learning, data mining and visualization techniques over the entire national healthcare data sets

Future work

- Privacy and security in the context of “Big Data” platforms
- Measuring and quantifying SQL/NoSQL architectural tradeoffs
- Research in advantages of graph-based and linked data (LoD)-based representation

Questions

Contact: Edmon Begoli

Oak Ridge National Laboratory, Oak Ridge, TN, USA

begolie@ornl.gov

865.241.1923