5 Tips for Building a Successful Big Data Strategy

Janet George, Chief Data Scientist, SanDisk

December 09, 2015

Forward-Looking Statements

During our meeting today we will make forward-looking statements.

Any statement that refers to expectations, projections or other characterizations of future events or circumstances is a forward-looking statement, including those relating to market growth, industry trends, future products, product performance and product capabilities. This presentation also contains forward-looking statements attributed to third parties, which reflect their projections as of the date of issuance.

Actual results may differ materially from those expressed in these forward-looking statements due to a number of risks and uncertainties, including the factors detailed under the caption “Risk Factors” and elsewhere in the documents we file from time to time with the SEC, including our annual and quarterly reports.

We undertake no obligation to update these forward-looking statements, which speak only as of the date hereof or as of the date of issuance by a third party, as the case may be.

Janet George
Chief Data Scientist

Big Data Platform/Data Science/ Cognitive Computing

BACKGROUND/RELEVANT EXPERIENCE/EXPERTISE

- SanDisk®: Build global core competencies. Shape, drive and implement the Big Data platform, products and technologies, using advanced analytics and pattern matching with semiconductor manufacturing data from the ground up. Industry experience, skillset and background are in Big Data Platform, Machine Learning, Distributed Computing, Compliers and Artificial Intelligence.
- Prior: Served as Managing Director/Chief Scientist/Big Data Expert at Accenture technology labs, responsible for Big Data Platform, Machine Learning, Cognitive Computing and open-innovation. Also served as Head of Yahoo Labs/Research Engineering inventing Next Generation Platforms, Cloud Infrastructures and Machine Learning for Big Data and also at eBay and Apple Computer amongst others.
- Education: Janet holds a Bachelors and Advanced Master Degree with distinction in Computer Science, Mathematics, with a thesis focus on Artificial Intelligence.

Agenda

- Big Data Challenges
- Common Mistakes
- Five Tips on Building a Successful Big Data Strategy
- Questions to Ask Yourself
- Q&A

Challenges to Building a Big Data Strategy

- UNITING the data for deep insights
  - Format compatibility (connective tissue) between completely different, variable data sets
- Preserving “DATA LINEAGE” for insightful analytics
- UNLOCKING DATA from traditional database systems
- TRANSCENDING Data Silos
- PLANNING for Data Growth
- MANAGING the Big Data Complexity

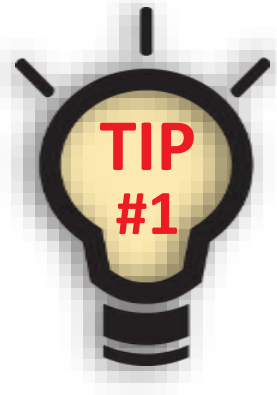
Common Mistakes

- ASSUMPTION: No need to unite data, work with small sample data sets
  - DANGER! Results often extrapolated to larger data sets; variance is not accounted for = misleading/skewed results
- ASSUMPTION: Advanced/complex algorithms (machine learning/data science) will solve all the problems
  - DANGER! Using uncleansed data to feed complex algorithms = garbage in garbage out

Five Tips to Building a Successful Big Data Strategy

SanDisk®
a Western Digital brand

Choosing the Right Enterprise Data Platform Strategy and Data Centric Architecture



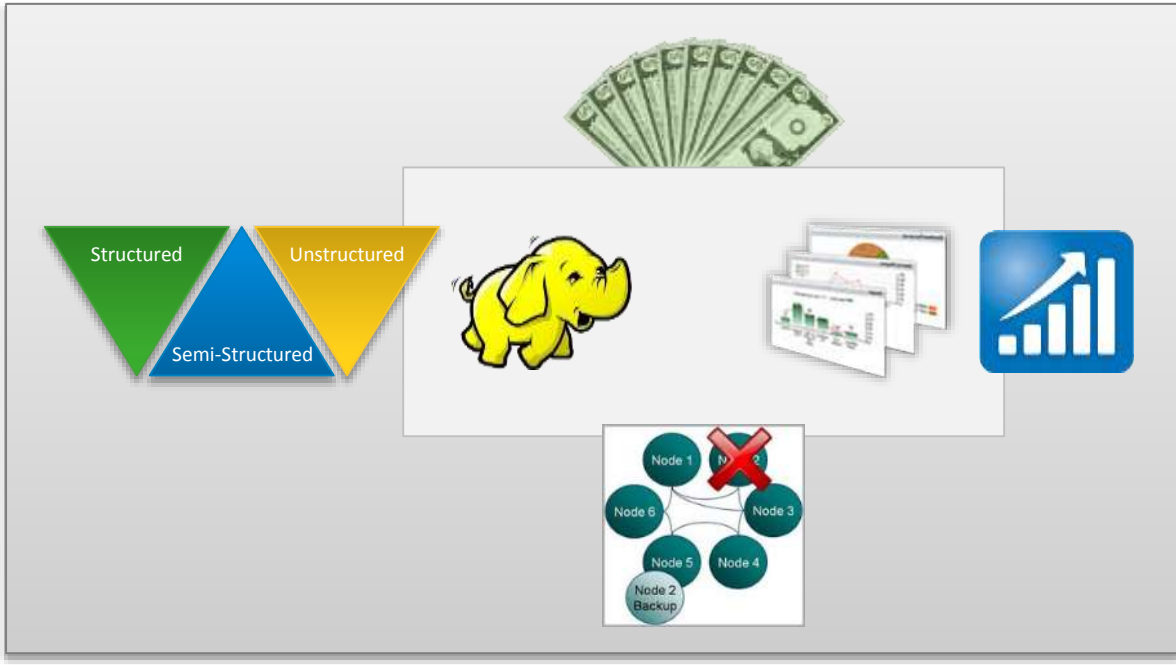
- Hadoop usual entry into the enterprise
- Enterprise Data Platform Strategy
- Hadoop distributions, MapR, Cloudera, Hortonworks? Which one to use? How to pick?
  - Efficient use of Hadoop with MapReduce for all batch processing, near real-time jobs running in background and doing parallel processing
  - Use spark heavily for all in-memory/large memory footprint processing within Hadoop stack

What's Hadoop?

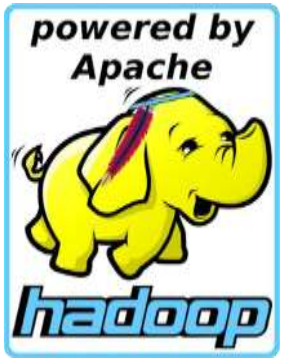
Apache Hadoop is an open source software project that enables distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines with a very high degree of fault tolerance.

Hadoop for Advanced Analytics

- Cost efficient for very large data sets
- Storage flexibility
- Scalability
- Self healing capability

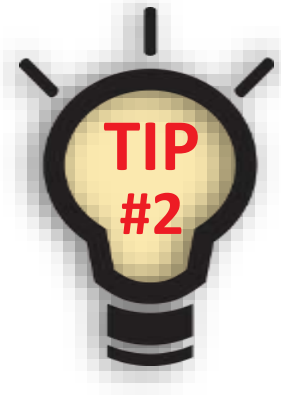


Hadoop: 10 Years Old and Enterprise Caliber



Operating System for the Data Platform – 4 Flavors

Embrace “Data Lake” and the Right Storage for Uniting all Your “Big Data”

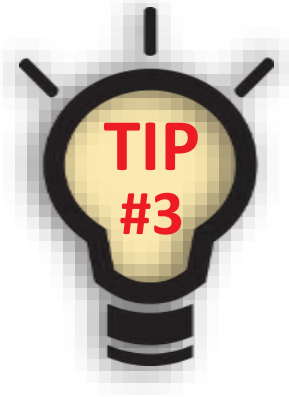


- Game changing and transformational in the journey towards connecting the enterprise data
- Highly scalable
- Massive capacity
- High performance
- Cost-effective
- Increased data availability and decreased effort in algorithm deployment
- Speed in data access

Data Lake

Dive into the Depths

Keep the Full Lineage of Data Wrangling and Cleansing within the "Data Lake"

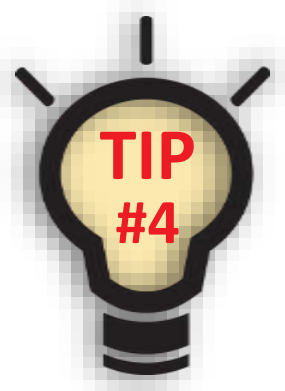


- Preserving full data lineage allows access to BOTH raw data and transformed data
- Storing in a data lake is far more cost effective than storing in Hadoop (three copies of data in Hadoop)
- Cataloged to find and retrieve data on an as-needed basis
- Ability to map data across sources and provide visibility to users
- Variable data sets, live seamlessly without conflict
- Very efficient tracking of cold, warm and hot data
- Data Lake becomes the default data destination with fine grained security and governance built in from the get go

Benefits of “Data Lake”

- Unified, central repository for all kinds and types of data, structured, unstructured and semi-structured data
- In addition to internal, external and partner data
- Rapid, automated ingest framework with low latency
- Ability to do 1st order, 2nd order and 3rd order advanced analytics seamlessly
- Raw data and transformed data – reporting (what happened?) and predictive (what will happen?) can work together efficiently

Work with as Much Data as Available (Large Data Sets)



- More data beats better algorithms!

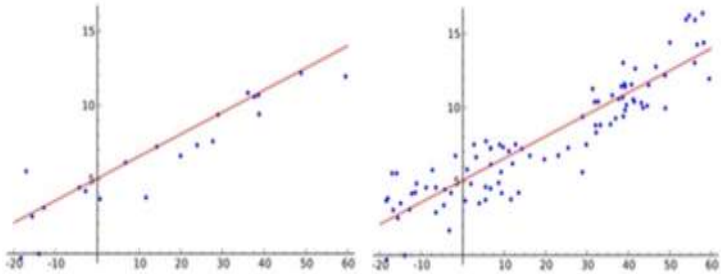


Figure 1. Estimating a linear relationship

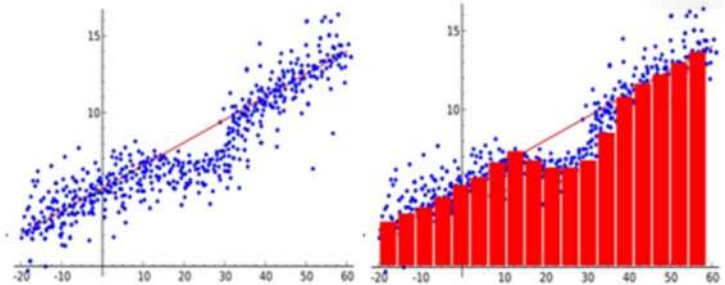
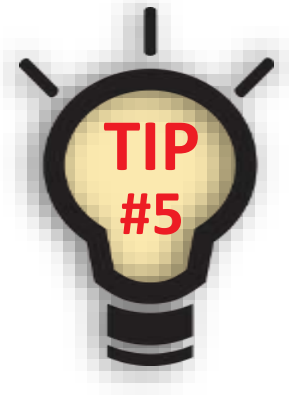


Figure 2. More data reveals a non-linear relationship in this dataset

- Reveals non-linear relationships, e.g. nonparametric models
- Allows building of more accurate models
- **Overall, a weak assumption coupled with complex algorithms is far less efficient than using more data with simpler algorithms**

Know Hadoop Limitations



- Small file issues
- Overcoming machine/source data limitations
- Maintaining positional context while combining files
- Variable schema, unification and restoration
- Disjoint columns in semi-structured data

Questions to Ask Yourself

- Do I have a “Big Data” Strategy that includes a data centric platform architecture?
- What is my strategy to unify my data from debilitating, entrenched siloes within my organization?
- What is my strategy to address growth of my data from all my existing disparate source and new sources without expensive duplication and redundant data storage?
- How will my organization mature to and embrace 1st order and 2nd order advanced analytics?
- What is my strategy to manage the complexity of my Big Data with growth?

Expanding The Possibilities of Storage

@SanDiskDataCtr

SanDisk Data Center Solutions

@BigDataFlash

itblog.sandisk.com