



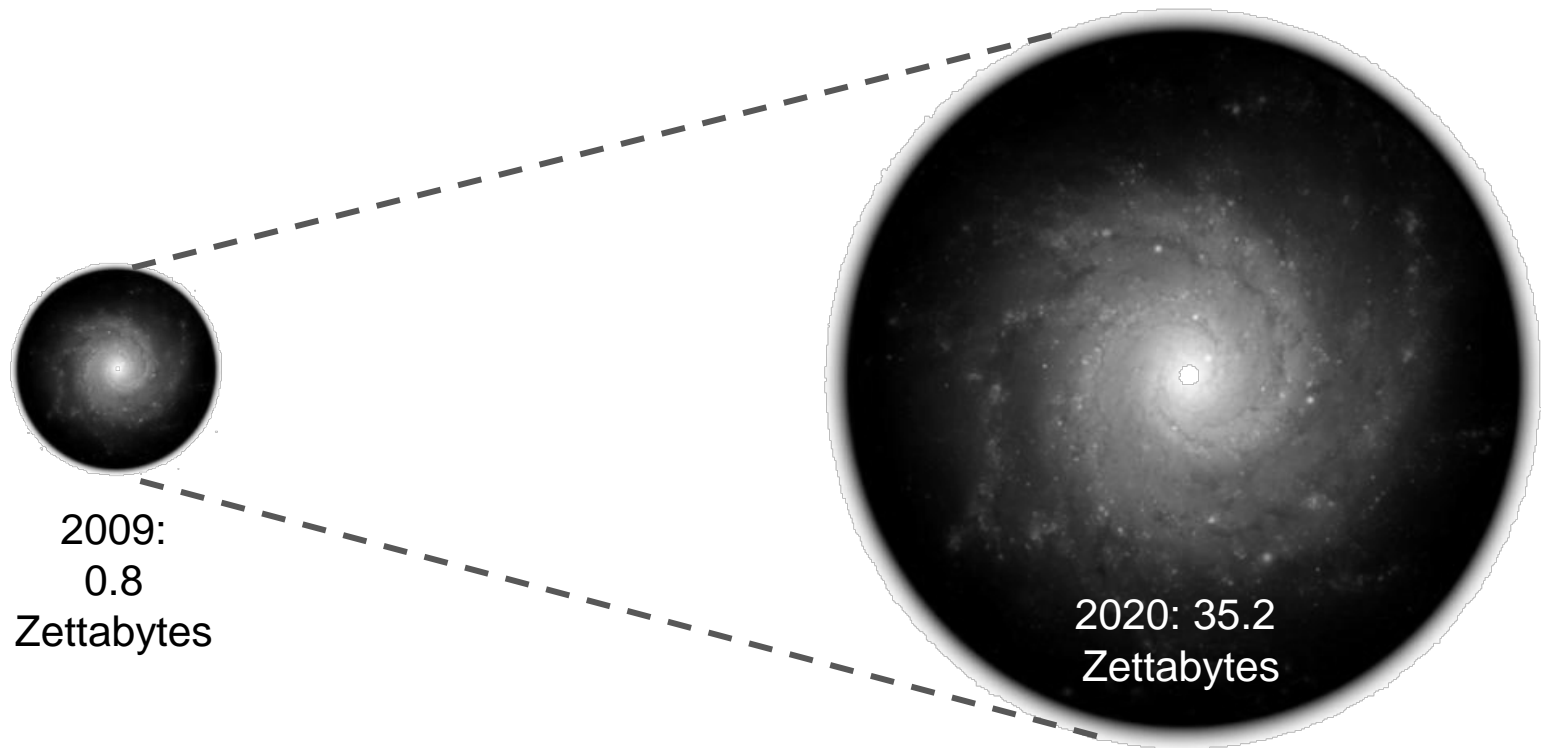
EMC Greenplum Driving the Future of Data Warehousing and Analytics

Tools and Technologies for Big Data

Steven Hillion
V.P. Analytics
EMC Data Computing Division

Big Data **Size**: The Volume Of Data Continues To Explode

The Digital Universe 2009 - 2020



Big Data Significance: Not Just For Google and Facebook...

“Just as search engines have transformed how we access information, other forms of *big data computing* can and will transform the activities of companies, scientific researchers, medical practitioners, and our nation's defense and intelligence operations.”

Randal E. Bryant

Carnegie Mellon
University

Randy H. Katz

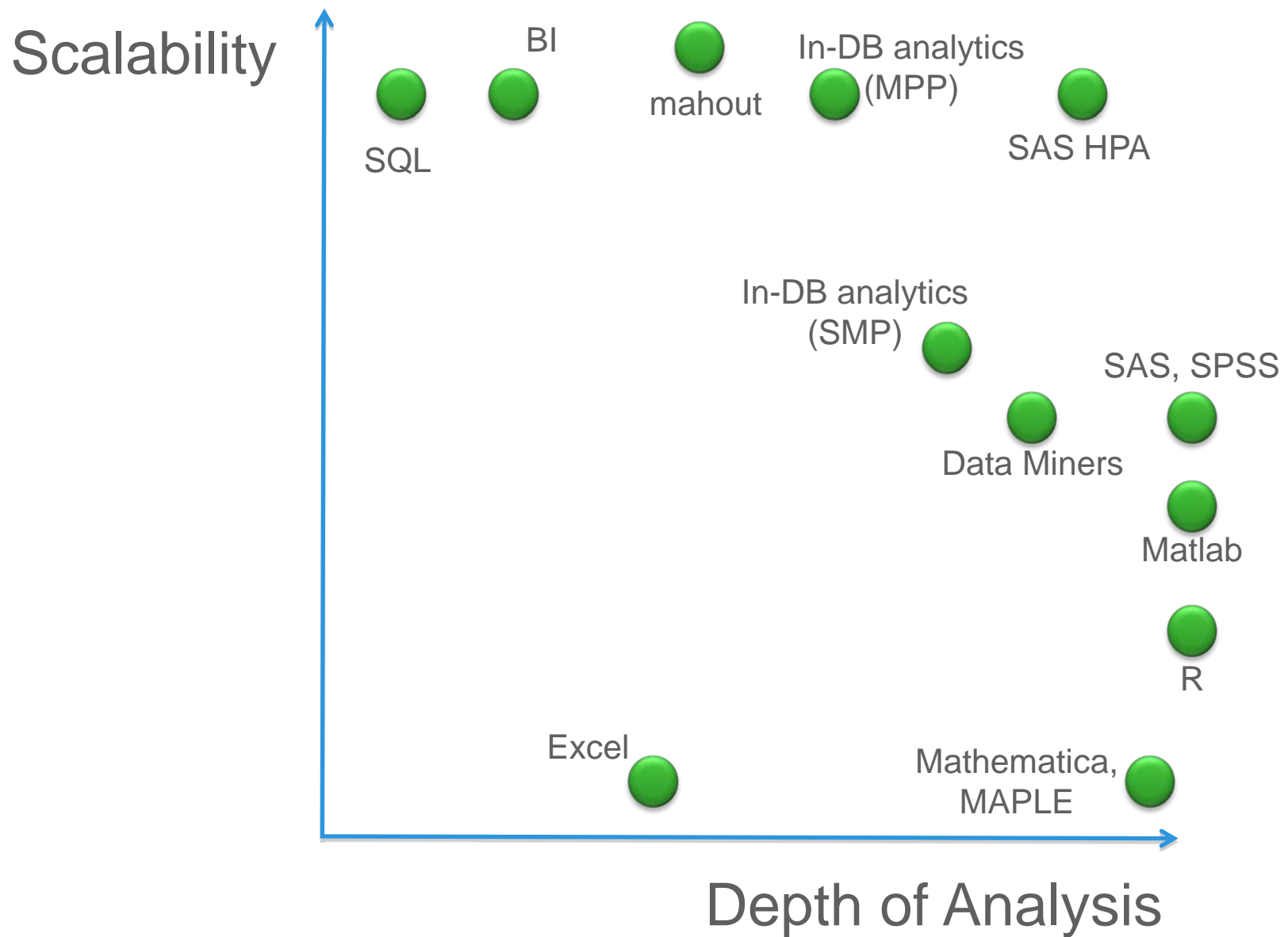
UC Berkeley

Edward D. Lazowska

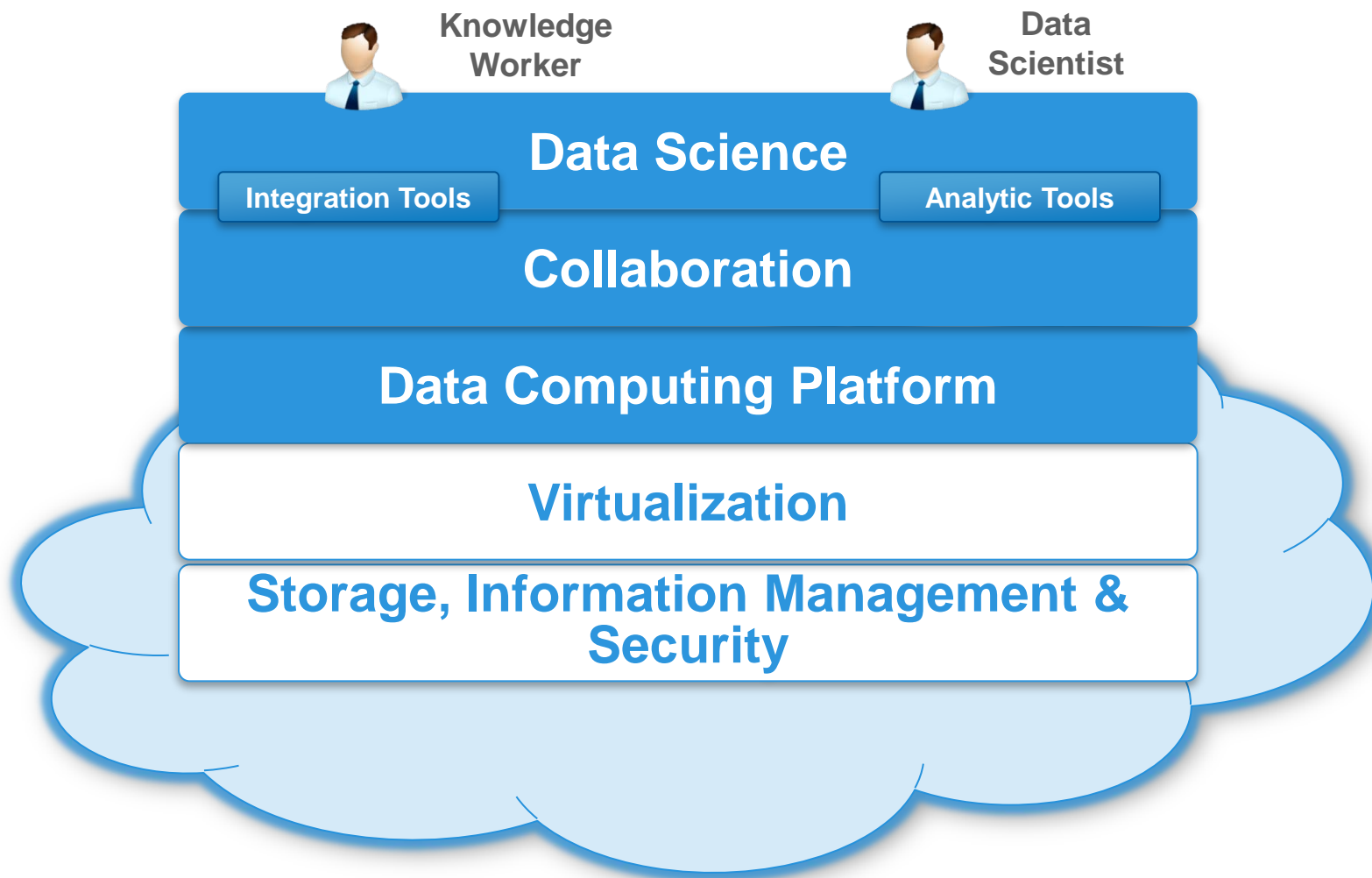
University of
Washington

Many Requirements, Many Technologies

- ETL for structured and unstructured data
 - Data storage
 - Computational infrastructure
 - Reporting and dashboards
 - Data mining
 - Model deployment
 - Visualization
 - Collaboration and Development
- e.g. MapReduce, ETL tools, SQL
 - e.g. HDFS, RDBMS
 - e.g. RDBMS, Hadoop, SAS Grid
 - e.g. Excel, BI Tools
 - e.g. SAS, SPSS, R, MADlib
 - e.g. PMML, In-DB scoring
 - e.g. Excel, Tableau
 - e.g. Wiki, Sharepoint



The Big Data Analytics Stack



Analytics Leadership



SAS/ACCESS

SAS Scoring Accelerator

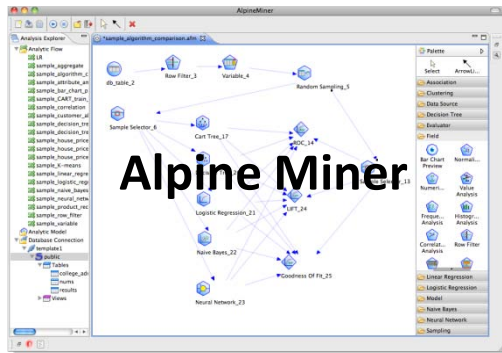
SAS High Performance Analytics

In-DB MapReduce
Greenplum HD

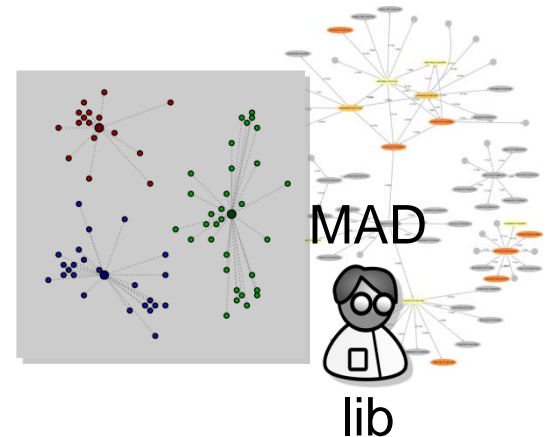


```
sql# SELECT cid, sum(tfxidf)/count(*) AS centroid
FROM (
  SELECT id, tfidf, cid,
  row_number() OVER (
    PARTITION BY id
    ORDER BY distance, cid) rank
  FROM blog_distance
) blog_rank
WHERE rank = 1
GROUP BY cid;
```

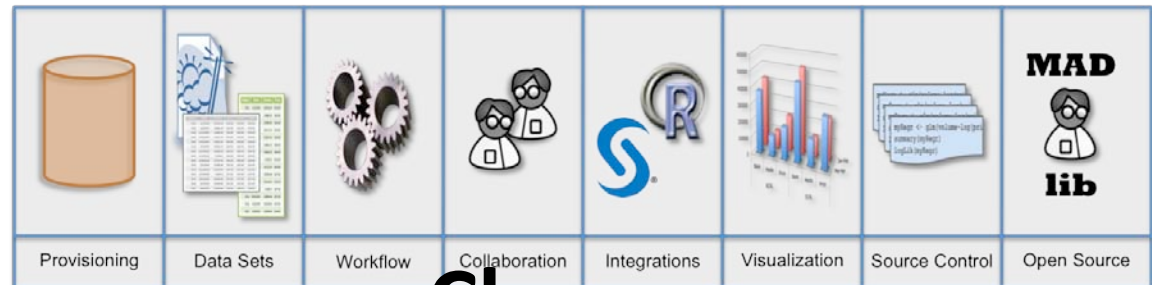
In-database Analytics



Alpine Miner



PL/R, Tools
integrations



Chorus

EMC²

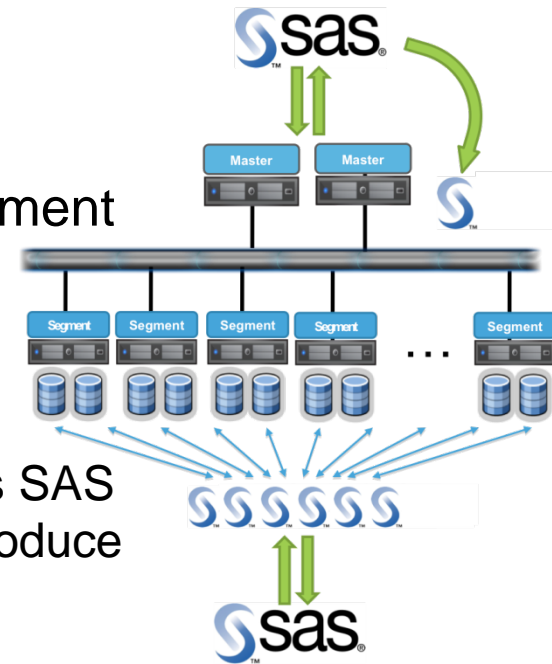
Main Page	Related Pages	Modules	Files	Search
<div> <div> ▼ MADlib <div> Main Page <div> Related Pages <div> Modules <div> Data Modeling <div> Supervised Learning <div> Naive Bayes Classification (Multi-)Linear Regression Logistic Regression Decision Tree Support Vector Machines <div> Unsupervised Learning <div> k-Means Clustering SVD Matrix Factorisation <div> Descriptive Statistics <div> Sketch-based Estimators <div> CountMin (Cormode-Muthukrishna) FM (Flajolet-Martin) MFV (Most Frequent Values) Profile Quantile <div> Support Modules <div> Sparse Vectors Conjugate Gradient </div> </div> </div> </div> </div> </div> </div> </div> </div> </div> </div> </div> </div> <div> Naive Bayes Classification </div> <div>Supervised Learning</div> <div> Collaboration diagram for Naive Bayes Classification: </div> <div> About: </div> <div> Naive Bayes classification with user-defined smoothing factor (default: Laplacian smoothing). </div> <div> A Naive Bayes classifier computes the following formula: </div> <div> $\text{classify}(a_1, \dots, a_n) = \arg \max_c \left\{ P(C = c) \cdot \prod_{i=1}^n P(A_i = a_i C = c) \right\}$ </div> <div> where c ranges over all classes in the training data and probabilities are estimated with relative frequencies from the training set. </div> <div> There are different ways to estimate the feature probabilities $P(A_i = a C = c)$. The maximum likelihood estimate takes the relative frequencies. That is: </div> <div> $P(A_i = a C = c) = \frac{\#(c, i, a)}{\#c}$ </div> <div> where </div> <div> <ul style="list-style-type: none"> $\#(c, i, a)$ denotes the # of training samples where attribute i is a and class is c $\#c$ denotes the # of training samples where class is c. </div> <div> Since the maximum likelihood sometimes results in estimates of 0, it might be desirable to use a "smoothed" estimate. Intuitively, one adds a number of "virtual" samples and assumes that these samples are evenly distributed among the values attribute i can assume (i.e., the set of all values observed for attribute a for any class): </div> <div> $P(A_i = a C = c) = \frac{\#(c, i, a) + s}{\#c + s \cdot \#i}$ </div> <div> where </div> </div></div>				

- MADlib is an open-source library for scalable in-database analytics, jointly developed by EMC and UC Berkeley. It provides data-parallel implementations of mathematical, statistical and machine learning methods for structured and unstructured data.
- The MADlib mission: to foster widespread development of scalable analytic skills, by harnessing efforts from commercial practice, academic research, and open-source development.



A Strategic Partnership for High-Performance Computing and MAD Analytics

- Access relational data-sets for agile analysis
 - **SAS/ACCESS** provides fast, transparent and secure access to Greenplum data.
- Leverage database scalability for rapid model deployment
 - **SAS Scoring Accelerator** publishes models for execution in parallel across the Greenplum cluster.
- Build complex models at massive scales
 - The **SAS/Greenplum Analytics Appliance** combines SAS In-Memory Analytics with Greenplum parallelism to produce record-breaking scalability and performance.



Large-Scale, Real-World Analytics

Question	Method
How can I identify fraudulent activity?	Variable Selection, Logistic Regression
How do I segment my customers?	K-means Clustering
Does this product appeal to some segments more than others?	Log-likelihood
Which campaign is working better?	Mann-Whitney U Test
How is product ownership distributed across customer segments?	SQL, Cumulative Distribution Functions
How do I target my marketing efforts towards customers most likely to churn?	Logistic Regression
What new products should I offer my customers?	Cosine similarity, k-Nearest Neighbors
What are my customers saying about the new product launch?	MapReduce, NLP, sparse vectors



Q&A