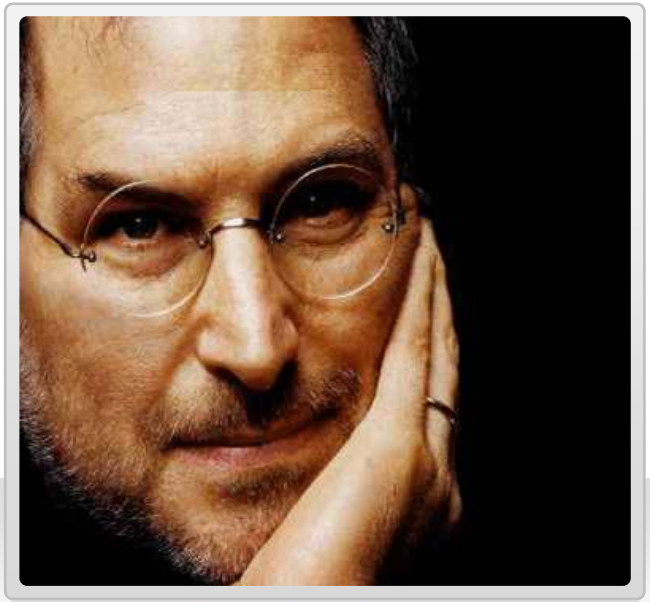
大数据，大机遇

陶波博士
EMC中国研发中心首席技术官
EMC中国研究院院长

EMC²



要实现阶段性变革、革命性变革，需要时机、技术、人才…的独特融合，以及对我们行业进行重大变革的运气。这种事情不会常常发生。

Steve Jobs，1995 年

快速膨胀中的各种数据来源

Source : 2011 IDC Digital Universe Study

EMC²

巨量的数目和巨大的文档

Files In The
Digital Universe

500
Quadrillion

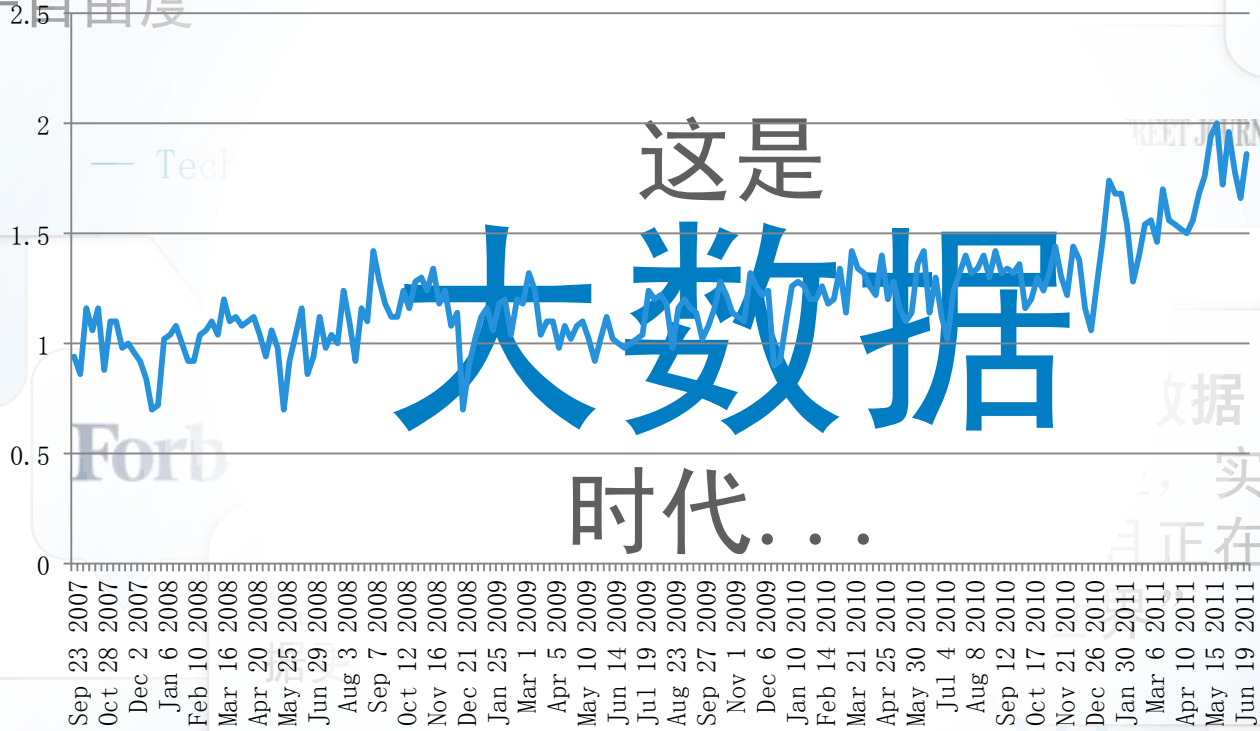
Big Data
Applications

5+ TB

Source : 2011 IDC Digital Universe Study, EMC Customers

“大数据无关乎大小，而关乎自由度”

谷歌趋势"big data"



这是

大数据

时代...

FORTUNE

New York Times

The Economist

— IDB

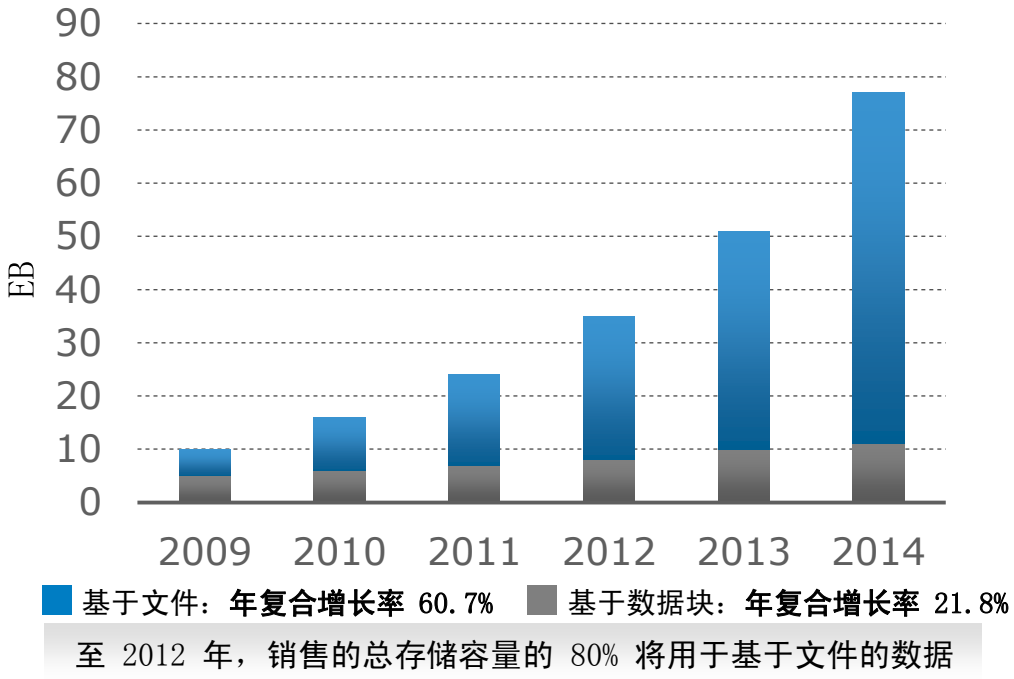
EMC 大数据“堆栈”

3 执行

2 实时/结构和非结构/协同

1 PB 规模

大数据正在改变企业存储



至 2012 年，销售的总存储容量的 80% 将用于基于文件的数据

来源：IDC

大数据要求：

容量和性能具有极大的可扩展性。

大数据系统需要适用巨量资料的存储架构

Scale Up, Manual	Scale Out, Automated
Storage Islands	One Storage Pool To 10+PB
More Capacity, More Admins	More Capacity, Same Admins
Performance Optimization “Whack-A-Mole”	Linear Performance Scalability

Isilon：横向扩展 NAS 创新

巨大的可扩展性

单个文件系统中超过 15 PB

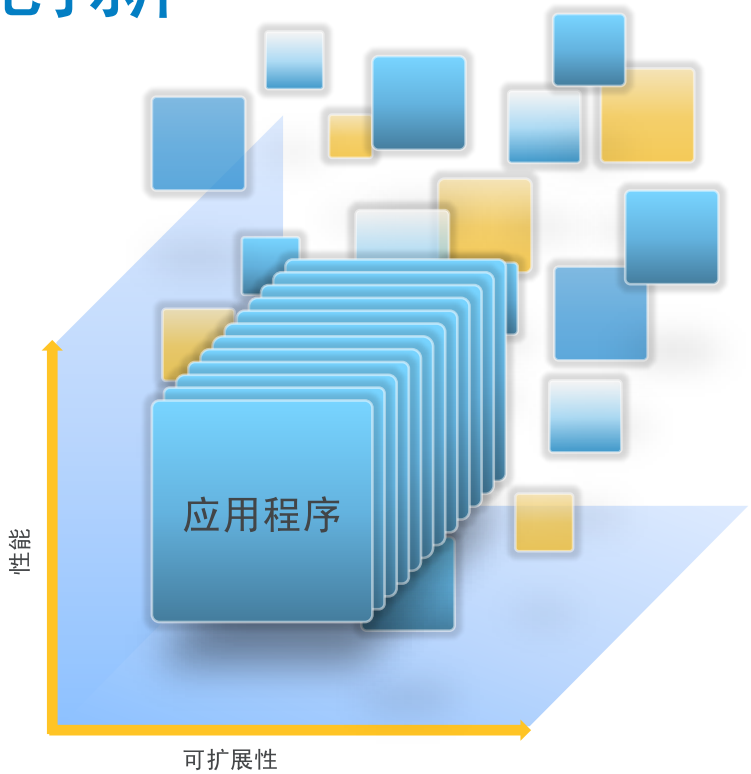
无可比拟的性能

高达 85 GB/s 的吞吐量和高于 1.2M 的 IOPS

应用程序与工作流整合

业界领先的可靠性和自我修复能力

管理简便



核心创新...为客户提供价值

Isilon 的 OneFS 横向扩展操作系统

- 单一文件系统，单卷...高达 15 PB 以上
- 原始存储利用率超过 80%
- 最高的性能，完全对称的群集
- 易于管理和扩展
- 多层单一文件系统/单群集
- 跨所有产品的单个统一平台

File Striping: Writing a File

Isilon 解决方案适用于...

企业 IT 扩展工作流

- 大规模主目录
- 大规模文件归档
- 灾难恢复与业务连续性

企业共享基础架构

- 私有云
- 第 3 层服务器虚拟化
- 存储整合

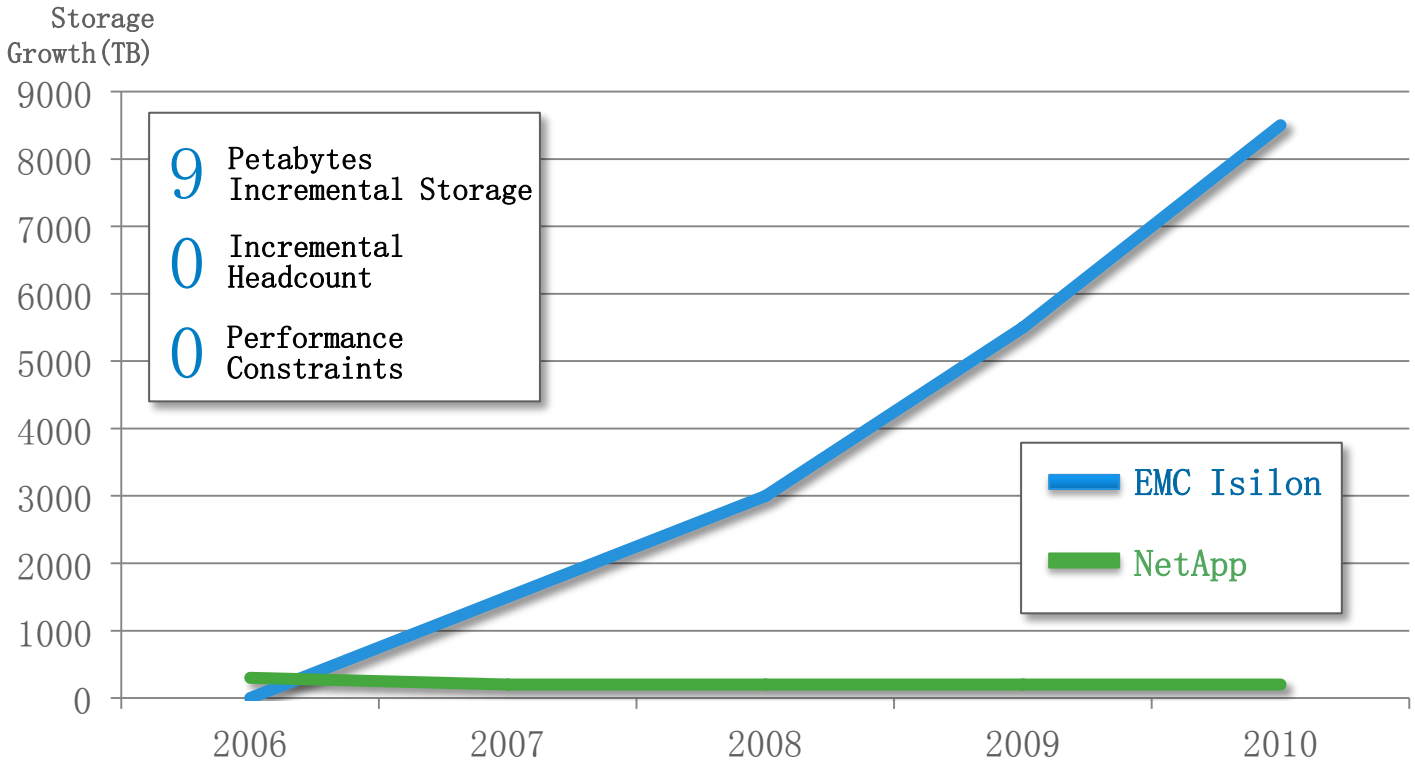
行业解决方案

- 媒体和娱乐
- 生命科学
- Internet 与 Web 2.0
- EDA 与软件开发

高性能计算

- 定量财务
- 地震处理
- 研究与分析
- 生命信息学

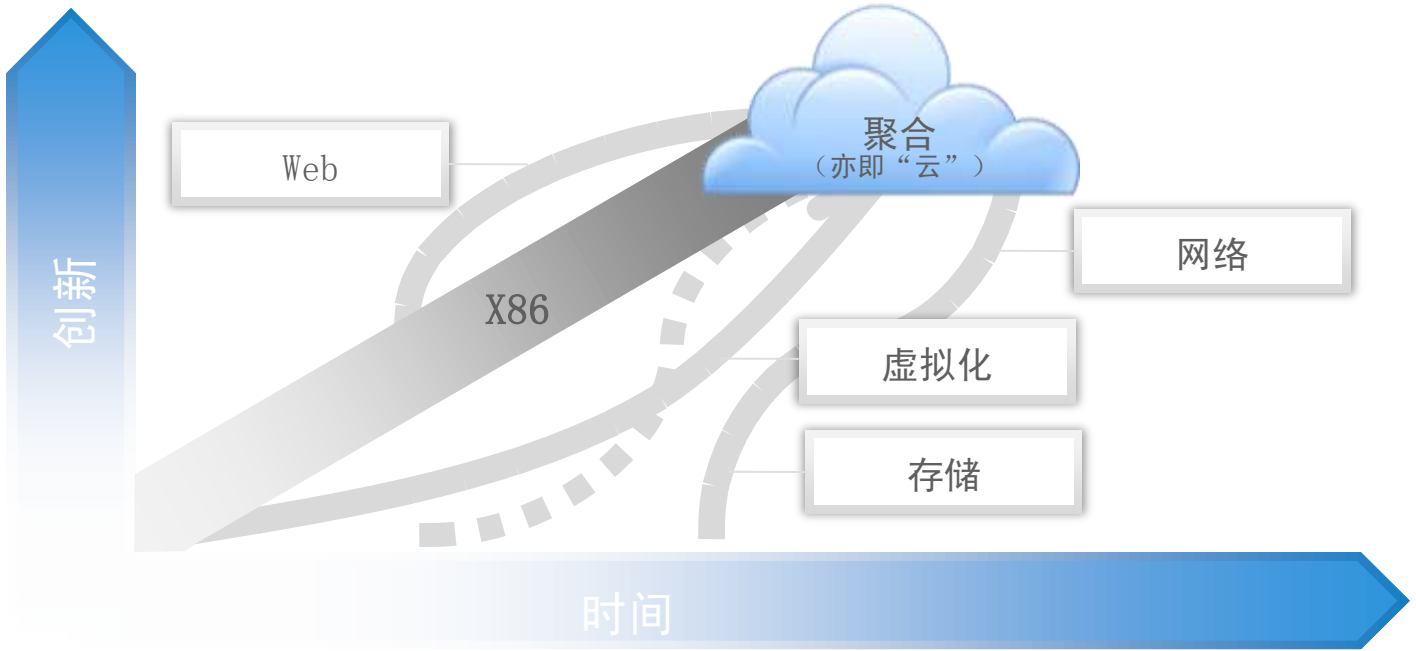
Isilon成功案例



”With Isilon, We’ve Been Able To Grow A Single File System To Nearly Nine Petabytes In Less Than Four Years, While Still Maintaining The Data Throughput Necessary To Power Our Workflow And Advance Our Research”

Matthew Trunnell
Manager, Research Computing Broad Institute

技术的聚合使大树据分析成为现实



我们还需要什么？

我们需要. . .

数据科学家

创新

社区

和

完整的大数据分析堆栈

WELCOME TO THE WORLD OF BIG DATA ANALYTICS

Structured. Unstructured. We've Got You Covered.

EMC GREENPLUM HD
Enterprise-Ready
Apache Hadoop

EMC GREENPLUM
DATABASE
Industry-Leading
MPP Performance

EMC GREENPLUM
COMMUNITY EDITIONS:
Greenplum HD
Greenplum Database

EMC GREENPLUM
CHORUS
The World's First
Enterprise Cloud Platform

EMC²

EMC HADOOP

非结构化。
实时。
企业就绪。

Greenplum HD 产品系列

- Greenplum HD **社区版**：
  - 经认证的满堆栈，100% 开源
  - 虚拟机装置
  - 所有核心功能开发反过来又有助于 Apache Hadoop
- Greenplum HD **企业版**：
  - 与众不同、混合分布、具有高级功能
  - 集成、经测试、加固型
  - 与 Hadoop、HBase、HDFS API 百分百兼容
- Greenplum HD Data Computing Appliance：
  - 优化装置配置
  - 消除复杂性，简化部署和管理
  - 与 Greenplum Database 无缝集成

EMC²

Greenplum HD 技术创新

可插拔 I/O

- Isilon OneFS
- Atmos
- Cassandra
- MapR
- 提高效率和性能

实时处理

- 低延迟读/写操作
- 实时数据交互和分析处理
- 与 Cassandra 和 MapR 集成

容错

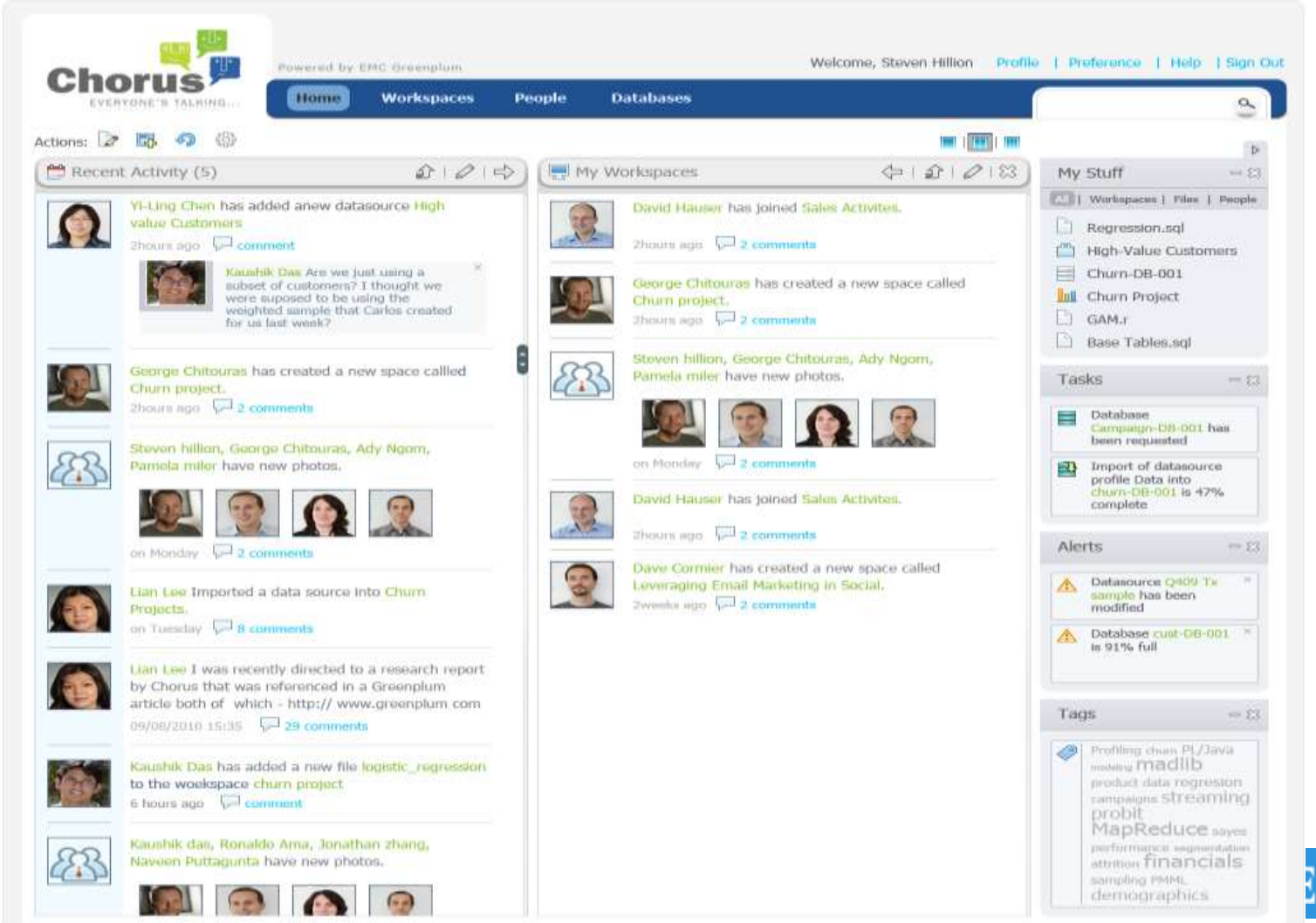
- 消除名称节点的单点故障
- 作业跟踪器及其他关键组件

GREENPLUM HD DATA COMPUTING APPLIANCE

Greenplum Database 与 Apache Hadoop 强强联合

Greenplum Chorus: 首创企业数据云平台

- 主要功能
  - *自我服务式生成数据库*
  - *数据服务*
  - *合作分析*
- Chorus 部署在VMware云计算平台和Greenplum数据库上
- Chorus极大地加速从数据中提取信息的过程


Chorus
EVERYONE'S TALKING...
Powered by EMC Greenplum
Welcome, Steven Hillion
Profile | Preference | Help | Sign Out
Home
Workspaces
People
Databases
Actions:
Recent Activity (5)
Yi-Ling Chen has added anew datasource High value Customers
2hours ago comment
Kaushik Das Ars we just using a subset of customers? I thought we were suposed to be using the weighted sample that Carlos created for us last week?
George Chitouras has created a new space callled Churn project.
2hours ago 2 comments
Steven hillion, George Chitouras, Ady Ngom, Pamela miler have new photos.
on Monday 2 comments
Lian Lee Imported a data source into Churn Projects.
on Tuesday 8 comments
Lian Lee I was recently directed to a research report by Chorus that was referenced in a Greenplum article both of which - http:// www.greenplum com
09/08/2010 15:35 29 comments
Kaushik Das has added a new file logistic_regression to the wookspace churn project
6 hours ago comment
Kaushik das, Ronaldo Ama, Jonathan zhang, Naveen Puttagunta have new photos.
My Workspaces
David Hauser has joined Sales Activites.
2hours ago 2 comments
George Chitouras has created a new space called Churn project.
2hours ago 2 comments
Steven hillion, George Chitouras, Ady Ngom, Pamela miler have new photos.
on Monday 2 comments
David Hauser has joined Sales Activites.
2hours ago 2 comments
Dave Cormier has created a new space called Leveraging Email Marketing in Social.
2weeks ago 2 comments
My Stuff
All | Workspaces | Files | People
Regression.sql
High-Value Customers
Churn-DB-001
Churn Project
GAM.r
Base Tables.sql
Tasks
Database Campaign-DB-001 has been requested
Import of datasource profile Data into churn-DB-001 is 47% complete
Alerts
Datasource Q409 Tx sample has been modified
Database cust-DB-001 is 91% full
Tags
Profiling chum PL/Java modeling madlib product data regreston campaigns streaming probit MapReduce sayes performance segmentation attrition financials sampling PMML demographics

自我服务可以快速开始一个新的项目

- **生成数据库服务器**，单节点或多节点
- **生成沙盒**用于分析.
- **方便地导入数据**.


Chorus
EVERYONE'S TALKING...
Powered by EMC Greenplum
Welcome, Steven Hillion
Profile | Preference | Help | Sign Out
Home
Workspaces
People
Databases
Actions:
Recent Activity (5)
My Workspaces
My Stuff
Alerts
Tags
创建合作环境对大数据做深度分析
○ 创立项目工作空间 共享文档，数据和工作流程.
○ 在沙盒中实施工作流程和管理相关变更.
○ 控制数据的权限.
○ 从in-database analytics functions中导入函数

EDC 成功实例：T-Mobile

100 TB EDW

1 Petabyte EDC

Customer Challenges

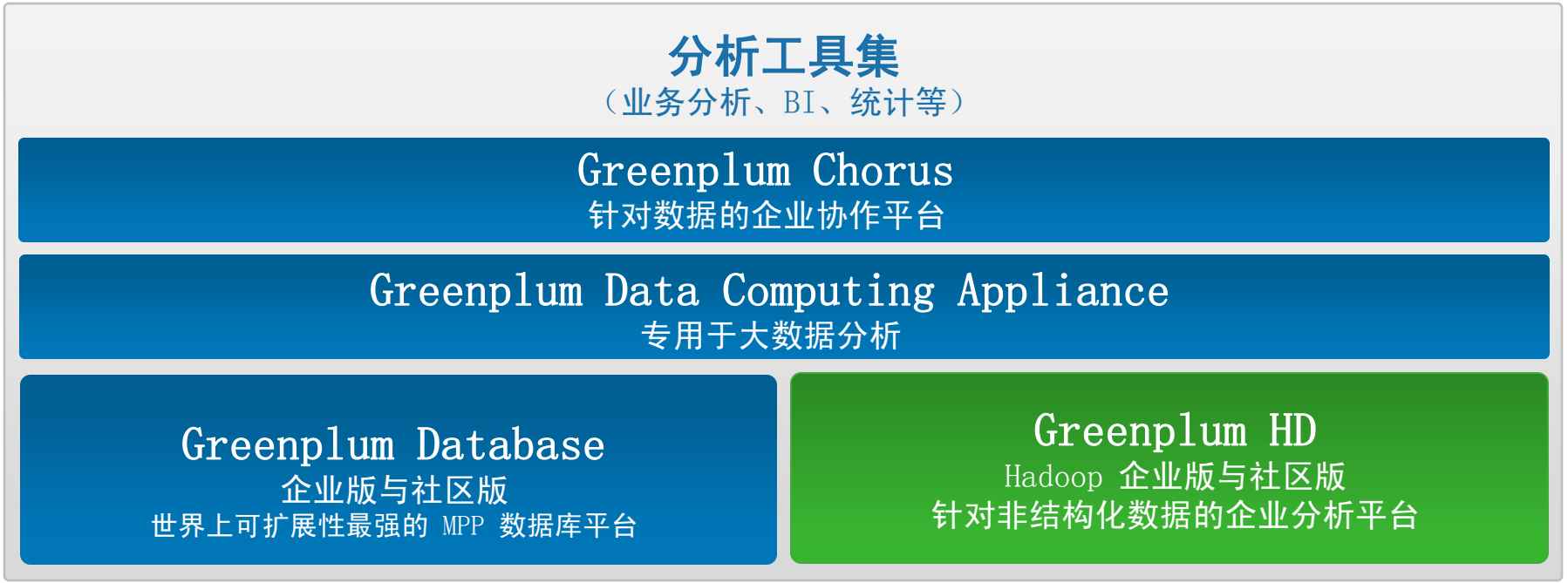
- 100TB Teradata EDW focused on operational reporting and financial consolidation
- EDW is single source of truth, under heavy governance and control
- Unable to support all of the critical initiatives around data surrounding the business
- Customer loyalty and churn the #1 business initiative from the CEO on down

EDC: Greenplum Database + Chorus

- Extracted data from EDW and others source systems to quickly assemble new analytic mart
- Generated a social graph from call detail records and subscriber data
- Within 2 weeks uncovered behavior where “connected” subscribers where 7X more likely to churn than average user
- Deployed1PB production EDC with GP to power their analytic initiatives

EMC²

构建完整的大数据分析堆栈



表彰大数据创新者

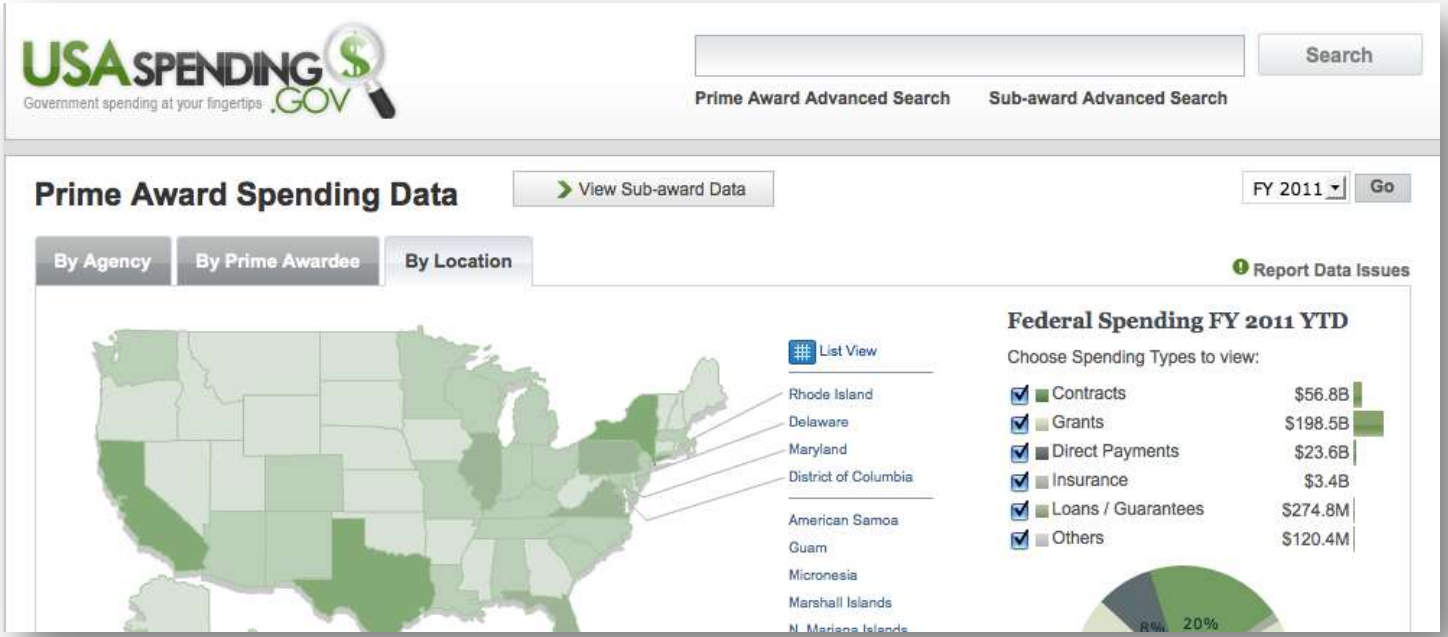
www.DataHeroAwards.com

“数据英雄奖”得主

Silver Spring Networks — 能源类

“数据英雄奖”得主

Vivek Kundra，美国首席信息官 — 远见奖



大数据 = 大机遇