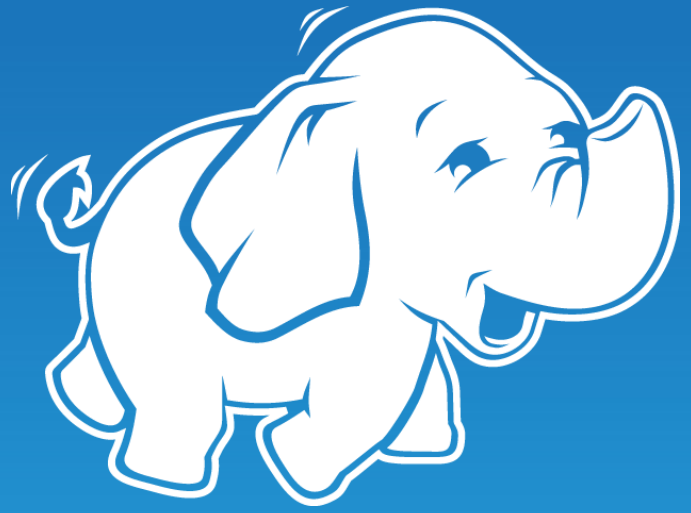
Windows Azure

大数据和Hadoop介绍



姓名
职位
公司

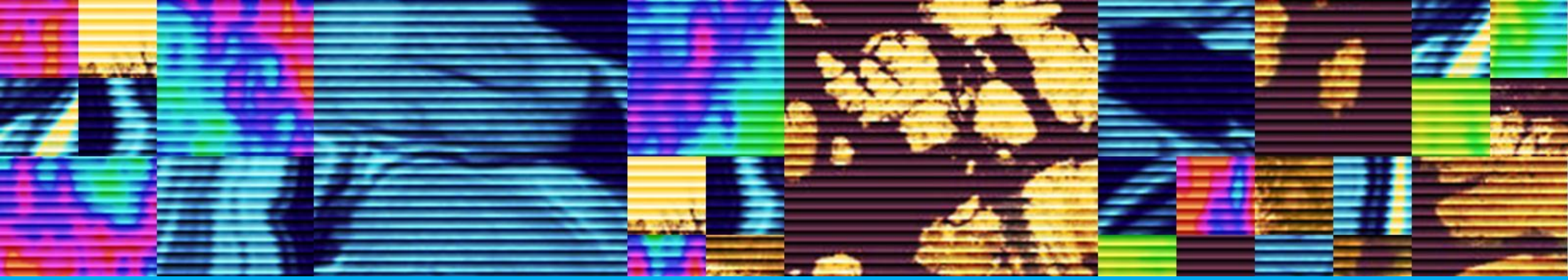
议程

什么是大数据?

理解基础

微软和Hadoop

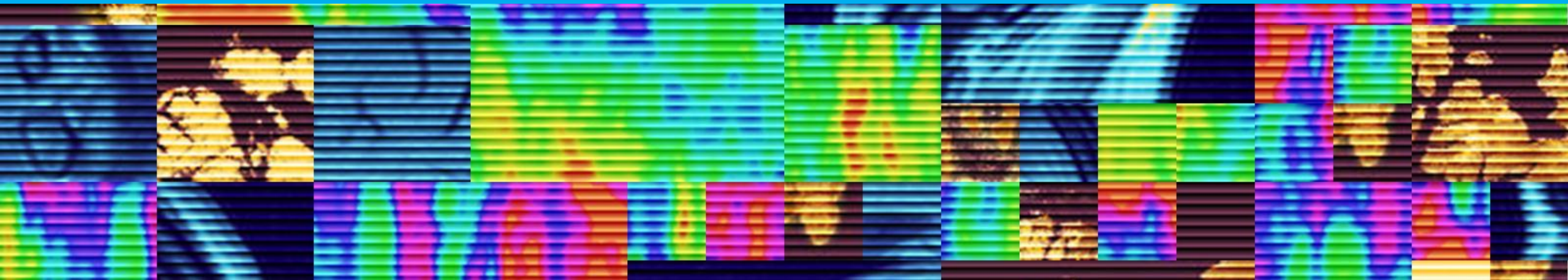
什么是大数据?

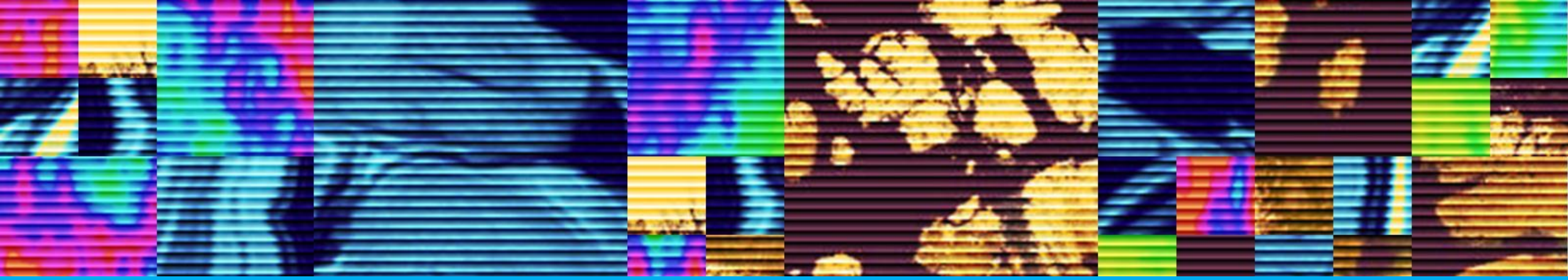


1.8 ZETTABYTES

信息将会在2011年创建

来源: CenturyLink 资源中心, 根据readwriteweb报告, 11月17日, 2011

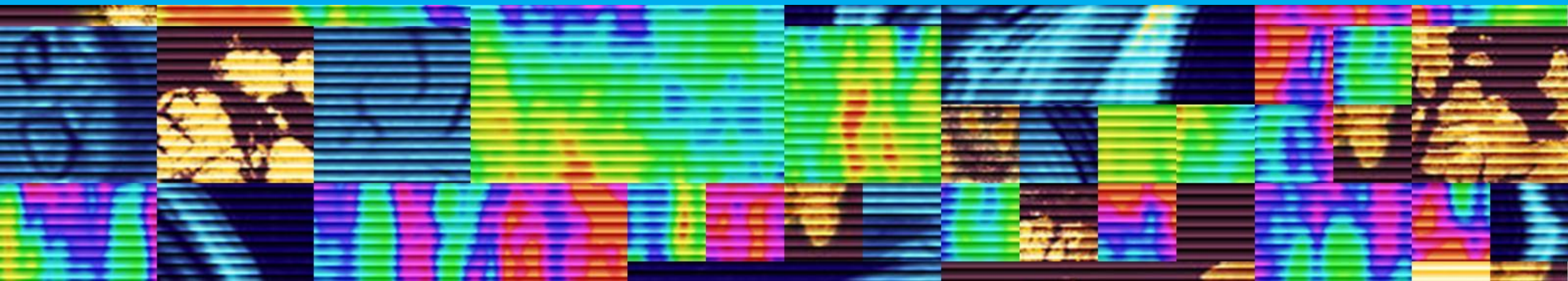




7.9 ZETTABYTES

2015前

来源: CenturyLink 资源中心, 根据readwriteweb, 11月17日, 2011报告



Bing 引入的数据> 7 petabytes 一个月

推特社区每天生成超过 1 terabyte 的推特信息

Cisco预测2013年每年互联网流量将会达到 667 exabytes

来源: 经济学人, 2010年2月; DBMS2; 微软公司

The Economist

FEBRUARY 27TH - MARCH 5TH 2010

Economist.com

Obama the warrior

Misgoverning Argentina

The economic shift from West to East

Genetically modified crops blossom

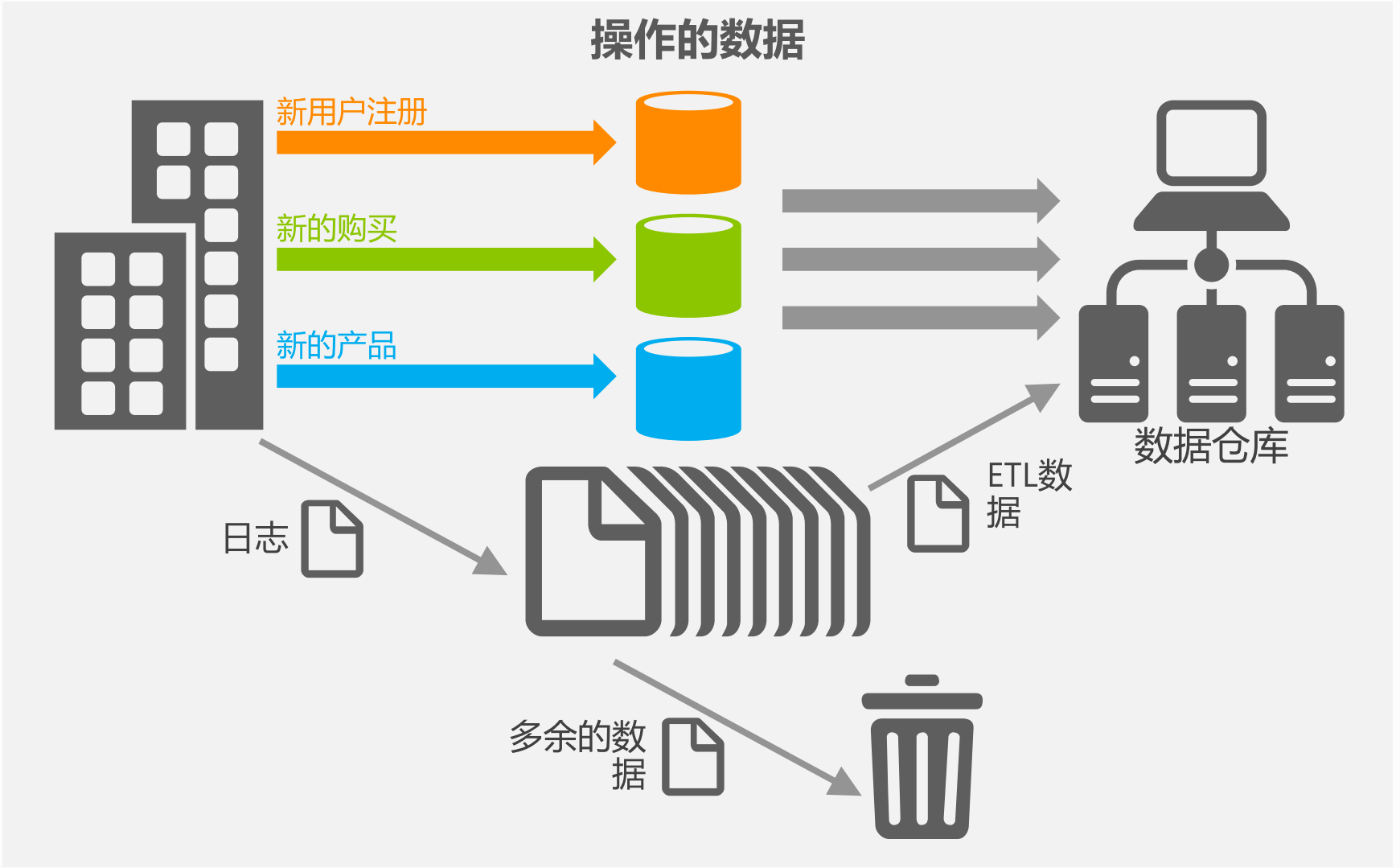
The right to eat cats and dogs

The data deluge

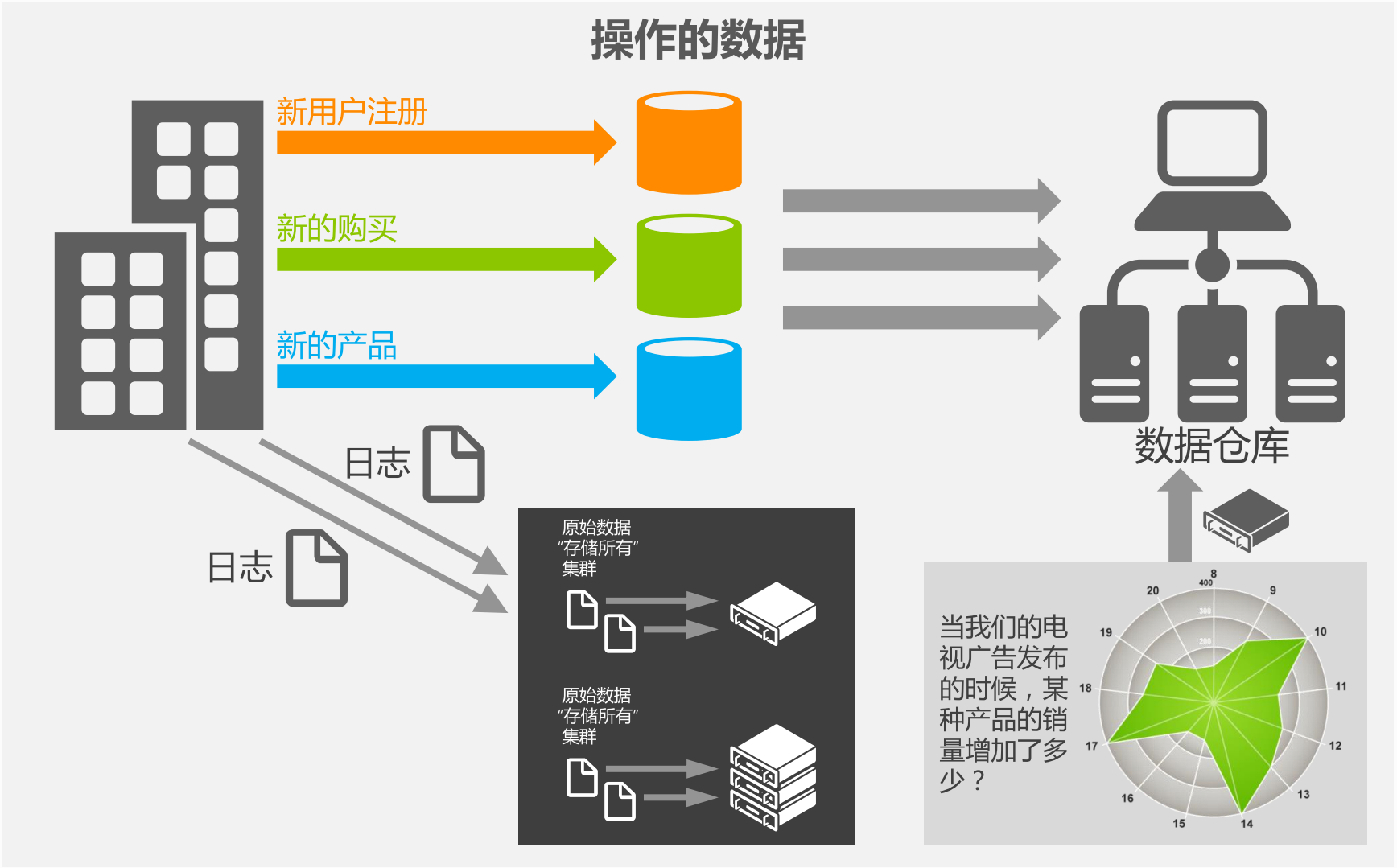
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

举例场景

传统电子商务数据流



新的电子商务数据流



Windows Azure

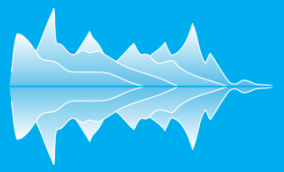
理解基础

把计算移向数据

Windows Azure™

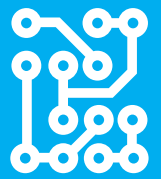
大数据的特点

新的数据源



非传统的数据类型

大数据量



新技术

新经济

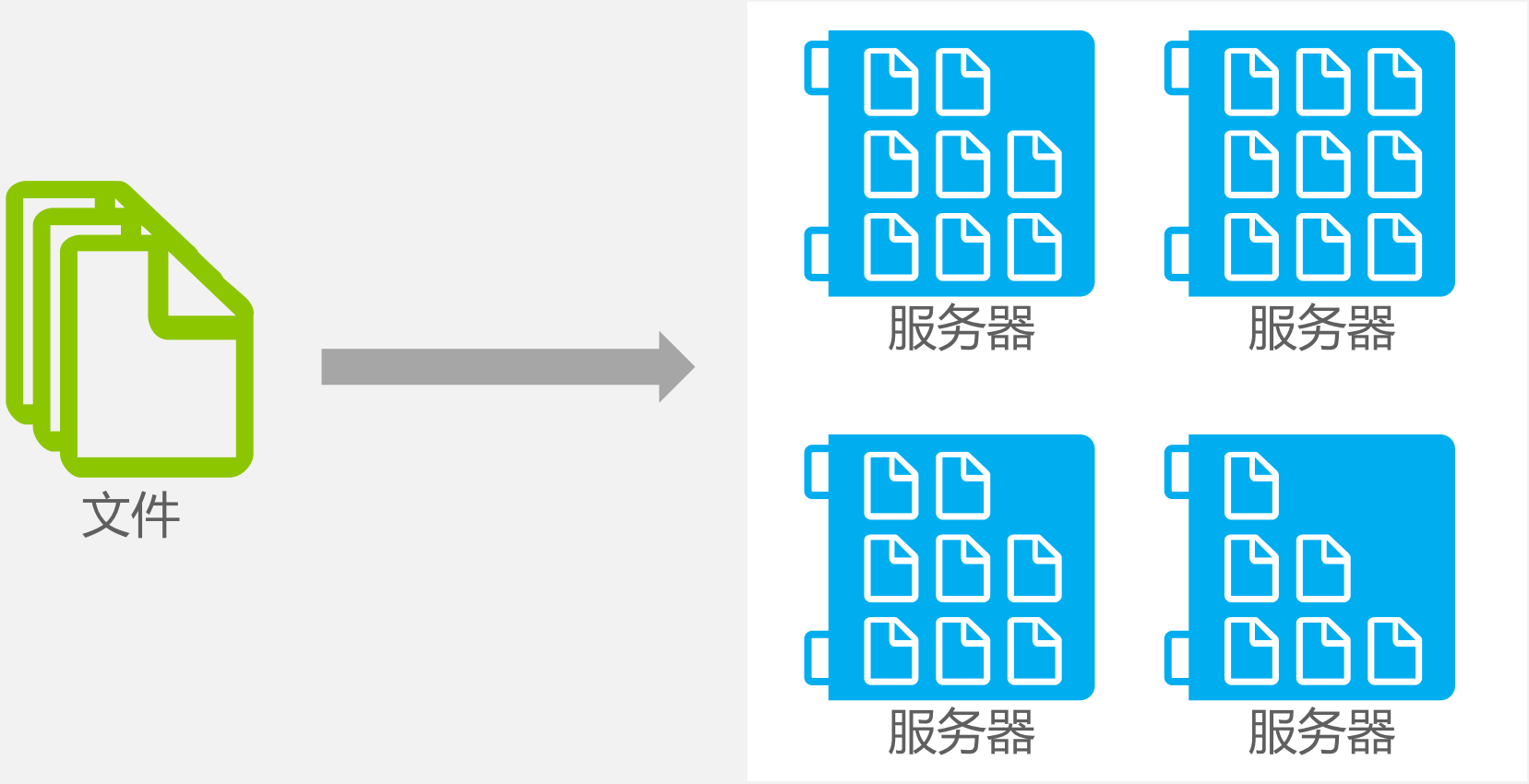
新的问题和新的视角

MapReduce

Windows Azure™

如何工作?

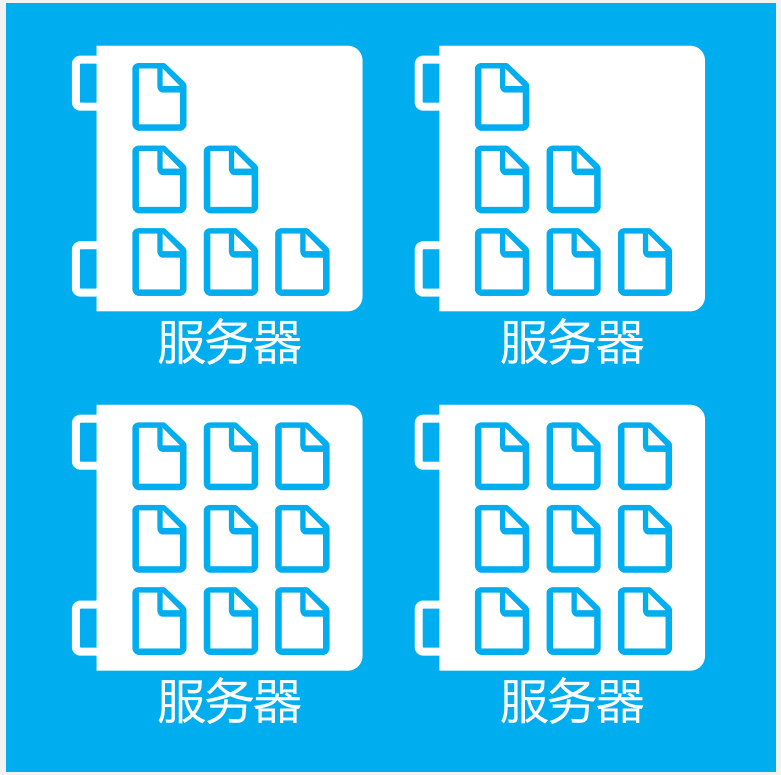
首先，存储数据



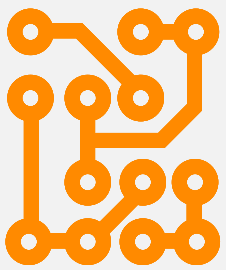
如何工作?

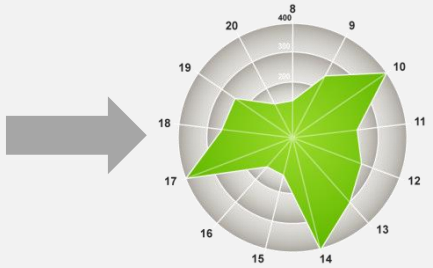
第二,处理数据

运行时



代码





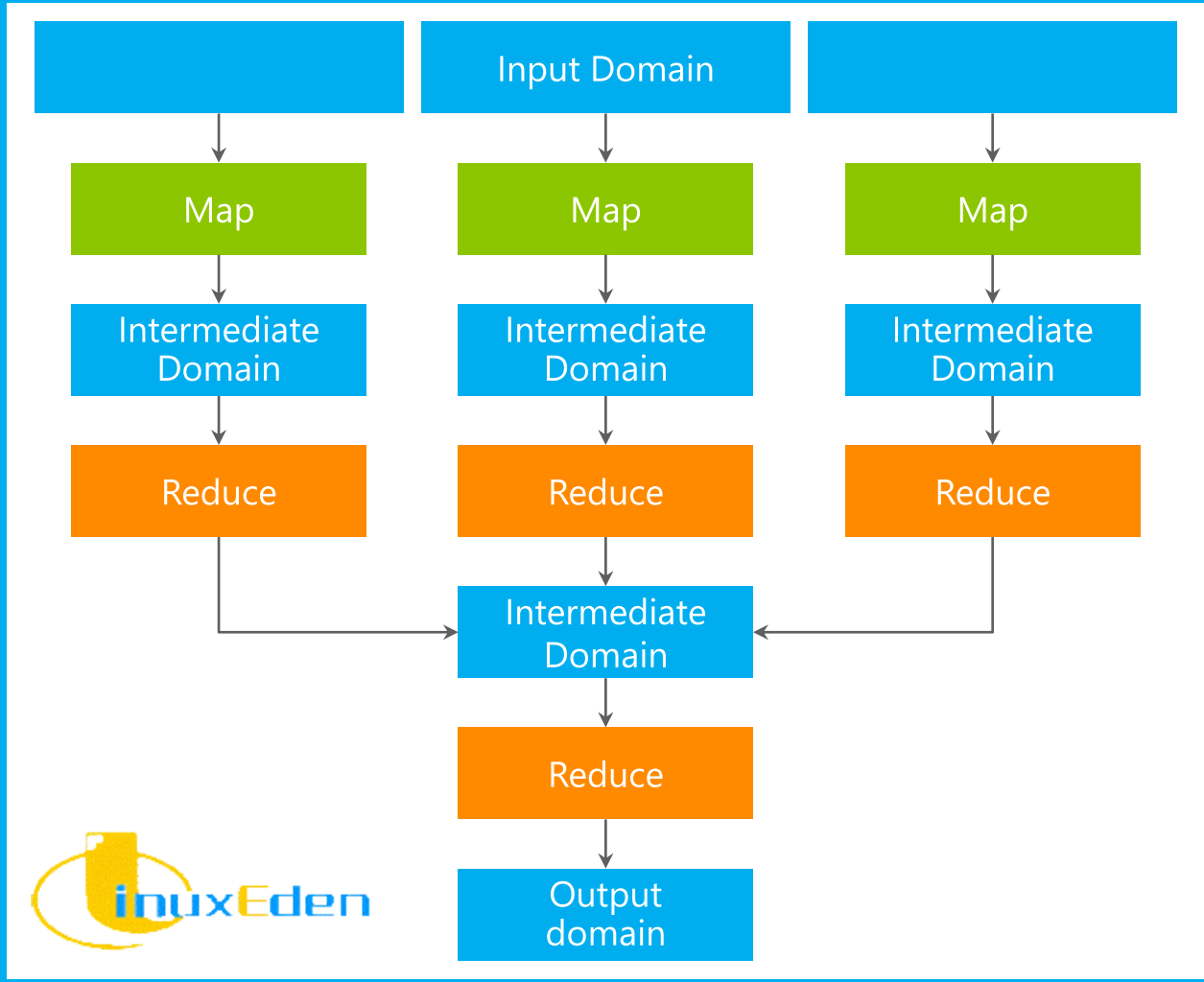
```
// Map Reduce function in JavaScript

var map = function (key, value, context) {
var words = value.split(/[^a-zA-Z]/);
for (var i = 0; i < words.length; i++) {
          if (words[i] !== "")
context.write(words[i].toLowerCase(),
1);}
}};

var reduce = function (key, values, context)
{
var sum = 0;
while (values.hasNext()) {
sum += parseInt(values.next());
     }
context.write(key, sum);
};
```

MapReduce – 工作流

一个MapReduce 工作通常把输入数据集分割成独立的块，这些独立的块用map任务以完全并行方式进行处理

框架把"map"的输出进行排序，然后输入到"reduce" 任务中

框架关心排程的任务，管理和重新执行失败的任务



MapReduce – 工作流

数据获取
和建模

合作
和可视化

分析
和数据挖掘

传播，分享
和保持

“It *takes more time to hand a project from the seismic guys to me to the engineers in production than it does to figure out the oil field plays.*”

Geologist,
Major oil and gas company

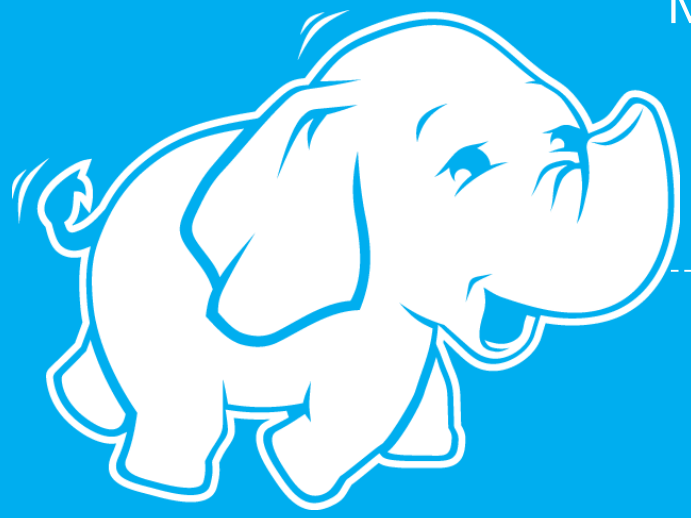
“*Our weather model and resulting data sets should be accessible to universities and other institutions.*”

Aerospace Development Manager,
U.S. Federal Government

Hadoop

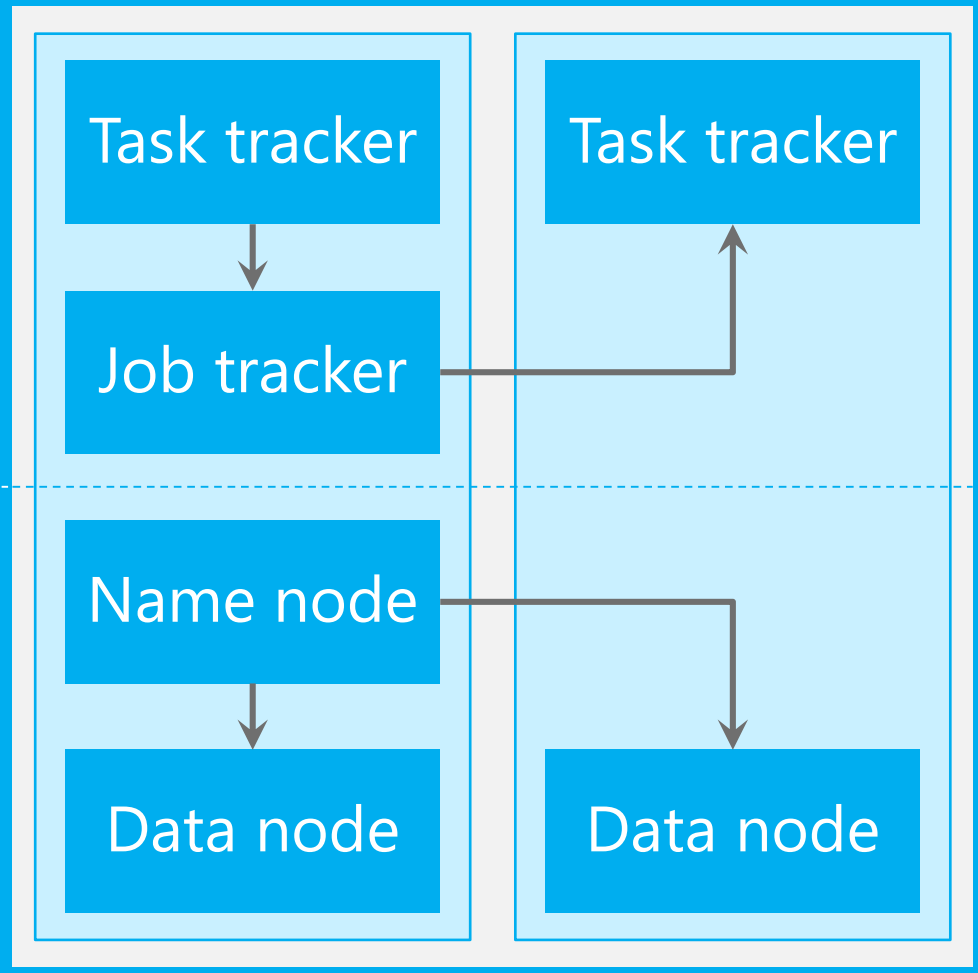
Windows Azure™

Hadoop 架构



MapReduce 层

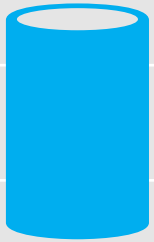
HDFS 层

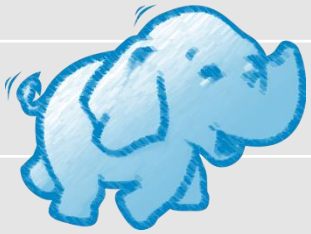


Reference: http://en.wikipedia.org/wiki/File:Hadoop_1.png

传统的RDBMS vs. MapReduce

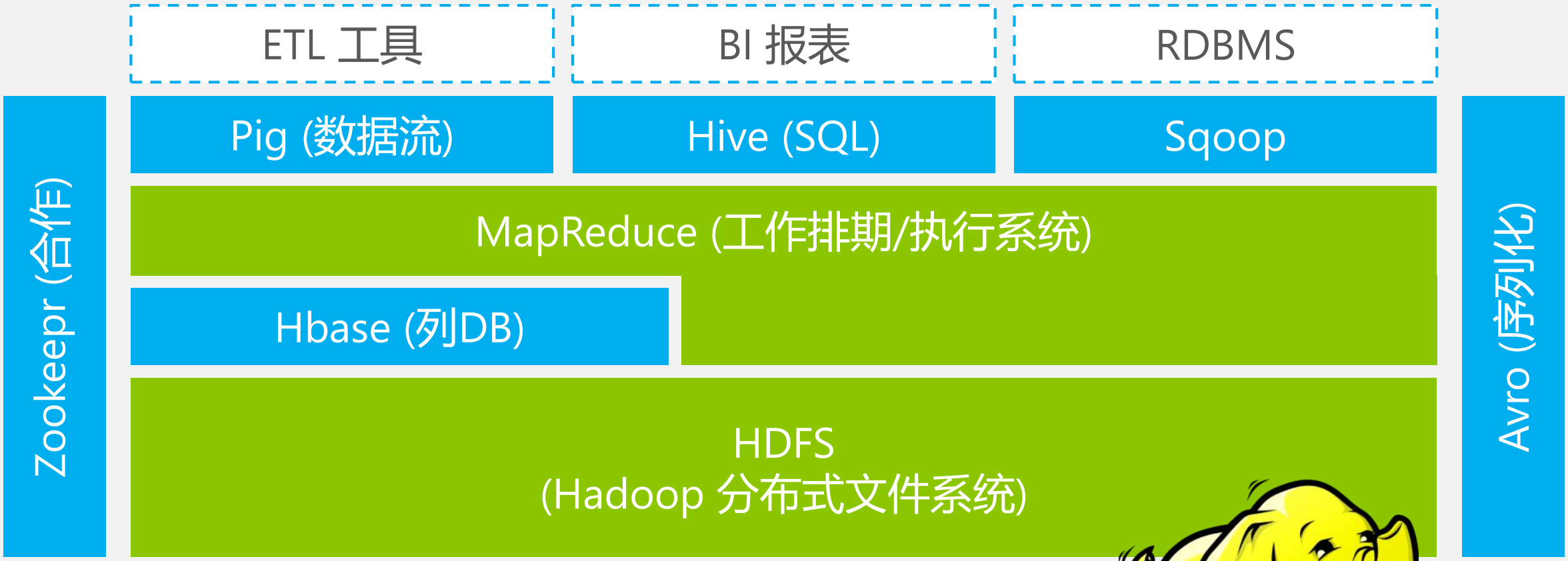
	传统RDBMS	MAPREDUCE
数据量	Gigabytes *(Terabytes)*	Petabytes *(Hexabytes)*
访问	交互和批量	批量
更新	多次读写	一次写，多次读
结构	静态结构	动态结构
数据完整性	高(ACID)	低
规模扩展	非线性	线性
DBA 比例	*1:40*	*1:3000*





Reference: Tom White's Hadoop: The Definitive Guide

The Hadoop 子系统



Reference: Tom White's Hadoop: The Definitive Guide

微软和Hadoop

Windows Azure™

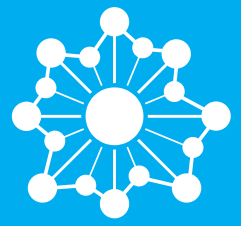
具体提供

深刻见解

Hive ODBC Driver 和 Hive Add-in for Excel
与Microsoft PowerPivot集成

企业就绪

Windows Server & Azure上基于Hadoop的分发
与Hortonworks战略合作



更广泛的访问

支持Hadoop的JavaScript框架
RTM of Hadoop connectors for SQL Server and PDW

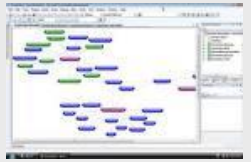
微软大数据解决方案

熟悉的最终用户工具

Power View

Excel和PowerPivot

预测分析



集成的BI

BI 平台

SSAS

SSRS

Hadoop On Windows Azure

Hadoop On Windows Server

连接器

Microsoft EDW

非结构化和结构化的数据

Sensors

Devices

Bots

Crawlers

ERP

CRM

LOB

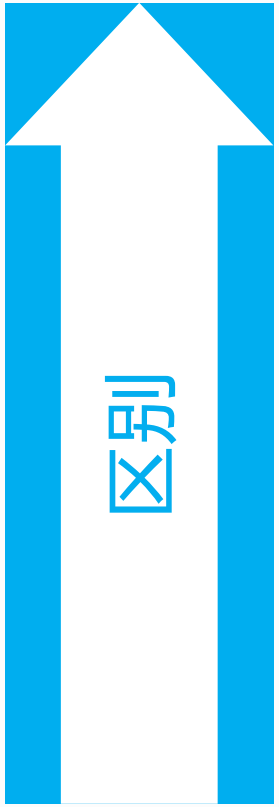
APPs

在Azure上部署和集成Hadoop集群

演示

Windows上的Hadoop

所有用户都能对新的数据类型有深刻见解

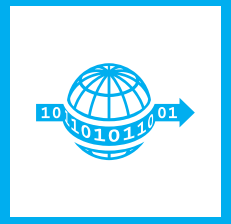


深刻见解	和微软商业智能集成
企业就绪	选择部署在Windows Server 和 Windows Azure上 和Windows 组件(AD, Systems Center)集成
更广泛的访问	在Windows上简单安装和配置 Hadoop 使用. Net和Javascript 集成简单编程 和SQL Server数据仓库集成

社区分发，贡献更多

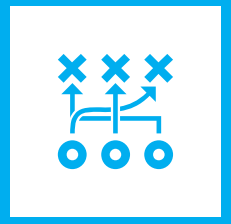
Windows Azure

微软大数据路线图

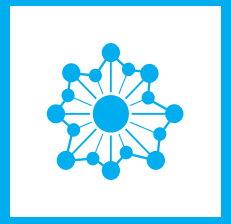


微软正在扩展他在商业智能和数据仓库的领导地位，把任何大小的数据的新的类型提供给所有用户

为了加速微软的基于Windows Server和Windows Azure的Hadoop解决方案，微软和Hortonworks成为合作伙伴



微软发布了大数据端对端的路线图， 包含了在Apache Hadoop™企业级的，在Windows Server和Windows Azure上的基于Hadoop的解决方案



微软承诺让所有规模组织的最终用户，开发人员和IT人员访问和使用Hadoop。

资源

http://www.hadooponazure.com/

http://hadoop.apache.org/

Microsoft®