

# Kapasiteettia kuin hanasta - tieteen pilvipalveluita: Microsoft Azure-palvelut tutkijoille

**Marko Hotti**

Senior Solution Architect

Microsoft Data Platform

[marko.hotti@microsoft.com](mailto:marko.hotti@microsoft.com)

<http://blogs.technet.com/markohot>

@MarkoHotti

# Microsoft end-to-end Data Platform from the functional point of view

Data Sources	Complex Event Processing	Enterprise Information Management	Enterprise Data Warehouse	Analytics Predictive Analytics and Data Mining	Reporting	Collaborative Platform	End User
 Oracle, Dynamics, Any LOB/data source		 SQL Server Integration Services	 SQL Server Parallel Data Warehouse	 SQL Server Analysis Services In-Memory Analytics	 SQL Server Reporting Services	 Office 365, Power BI Productivity and Collaboration	Office, Browser
		 SQL Server Master Data Services	 PolyBase (PDW) Single query model	 SQL Server Analysis Services Multidimensional	 Excel PowerPivot PowerView PowerQuery PowerMap PowerPoint SharePoint WebParts	 Office 365, Power BI, SharePoint Self-service BI and Analytics	Office, Browser
		 SQL Server Data Quality Services	 SQL Server PDW Hadoop Region	 SQL Server Analysis Services Datamining	 Office 365	 Office 365, SharePoint Reports, Dashboards and Scorecards	Office, Browser
 Sensor Data Log Data	 SQL Server StreamInsight		 WindoHDIInsight (ws Azure Hadoop)	 Azure CloudML Machine Learning		 Custom Applications	

# Cloud Computing



IaaS

Infrastructure-as-a-Service  
with Persistent remote disks

host



PaaS

Platform-as-a-Service  
Stateless, easy to scale, manage

build



SaaS

Software-as-a-Service

consume

# Why Cloud Computing?

- Abstraction for Hardware and IT Infrastructure
- Centralized Compute and data storage
- Convenience and Collaboration
- Always on and On Demand

# Compute and Optimizations

Design Optimization

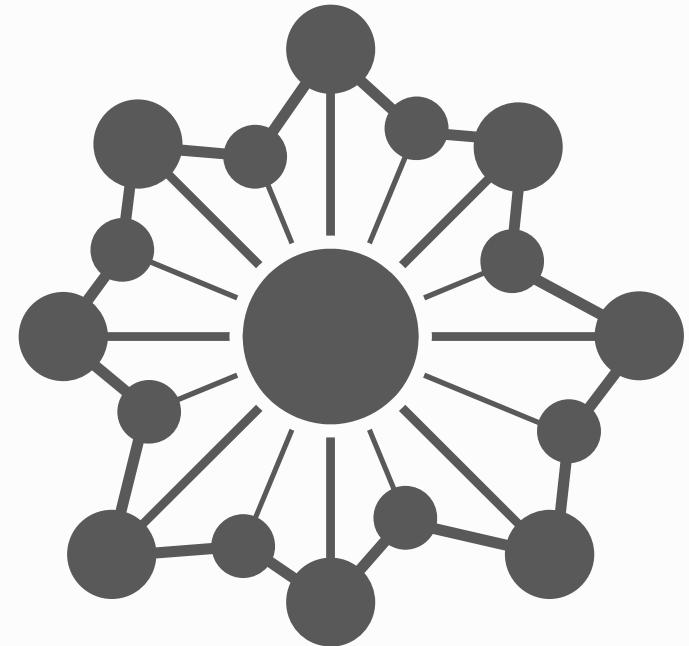
Financial and insurance risk calculation

Engineering modeling and simulation

Computational life sciences

Earth sciences

Data analytics



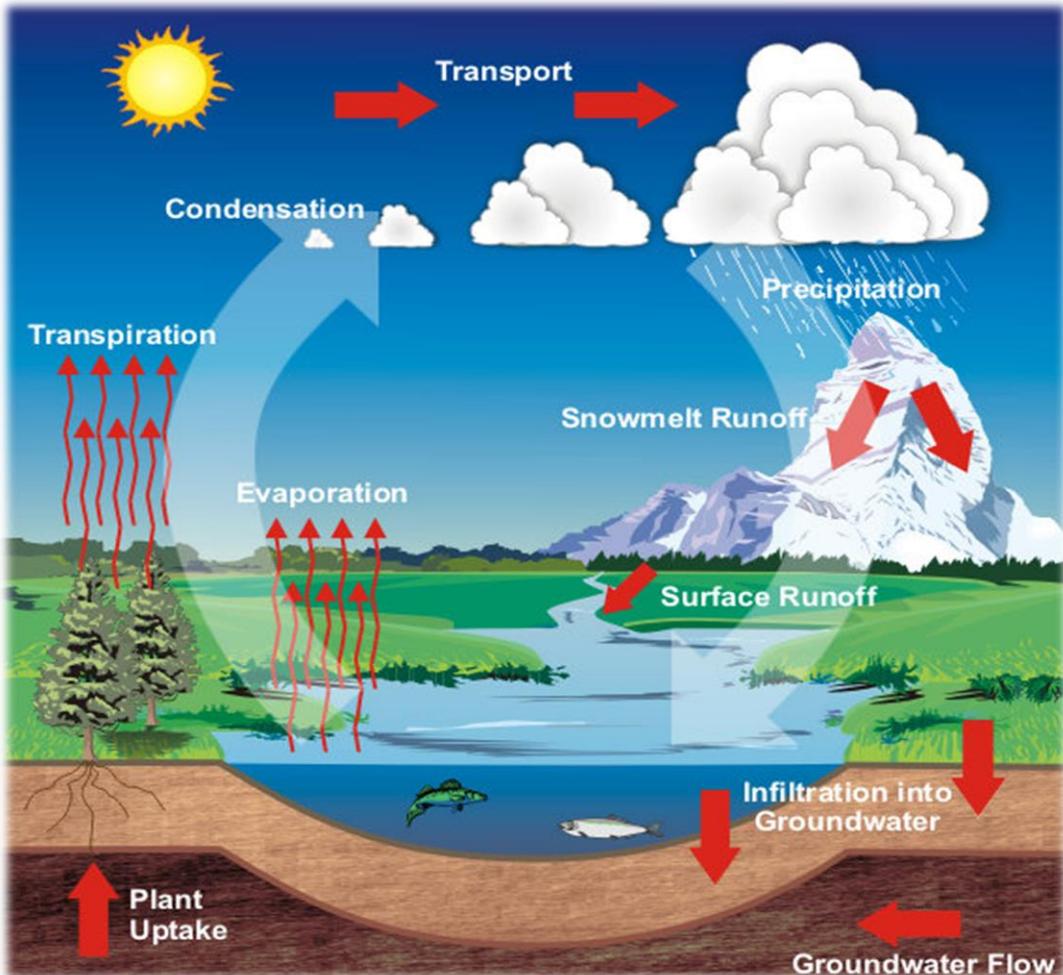
# Azure Scenarios in Scientific Research

- Scalable and cost-effective storage for scientific research data (data-at-rest, Hadoop processing, backups etc)
- HDInsight (100% Apache Compliant Hadoop Service in Microsoft Azure based on Hortonworks Data Platform)
- Virtual Machines (not just Windows Server but also Linux distributions) which can access Azure Storage
- Azure Machine Learning – a cloud service for running ML algorithms and R scripts against data in Azure
- Open Source applications, libraries, scientific research applications (Matlab etc) running on PaaS or IaaS services on Azure
- High Performance Computing (HPC) clusters
- Running Microsoft products and technologies in hybrid/cloud configurations
- Many more

Some real scientific scenarios and use cases we have done together with the academic world

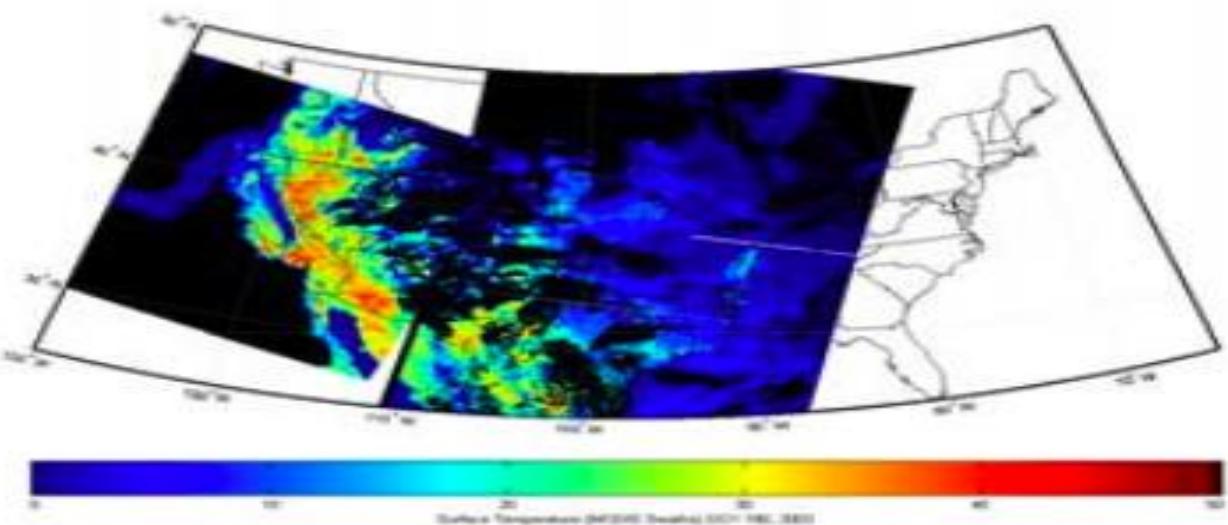
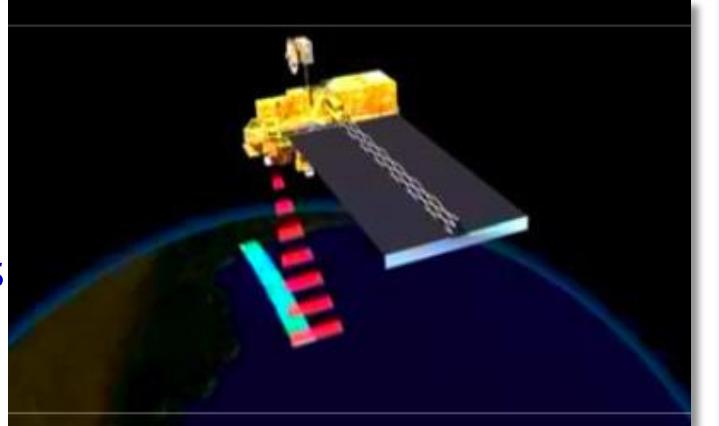
# AzureMODIS – Computing Evapotranspiration (ETP)

Catharine van Ingen (MSR), Jie Li, Marty Humphrey (UVA), Youngryel Ryu (UCB), Deb Agarwal (BWC/LBL), Keith Jackson (BL), Jay Borenstein (Stanford) , Team SICT: Vlad Andrei, Klaus Ganser, Samir Selman, Nandita Prabhu (Stanford), Team Nimbus: David Li, Sudarshan Rangarajan, Shantanu Kurhekhar, Riddhi Mittal (Stanford)

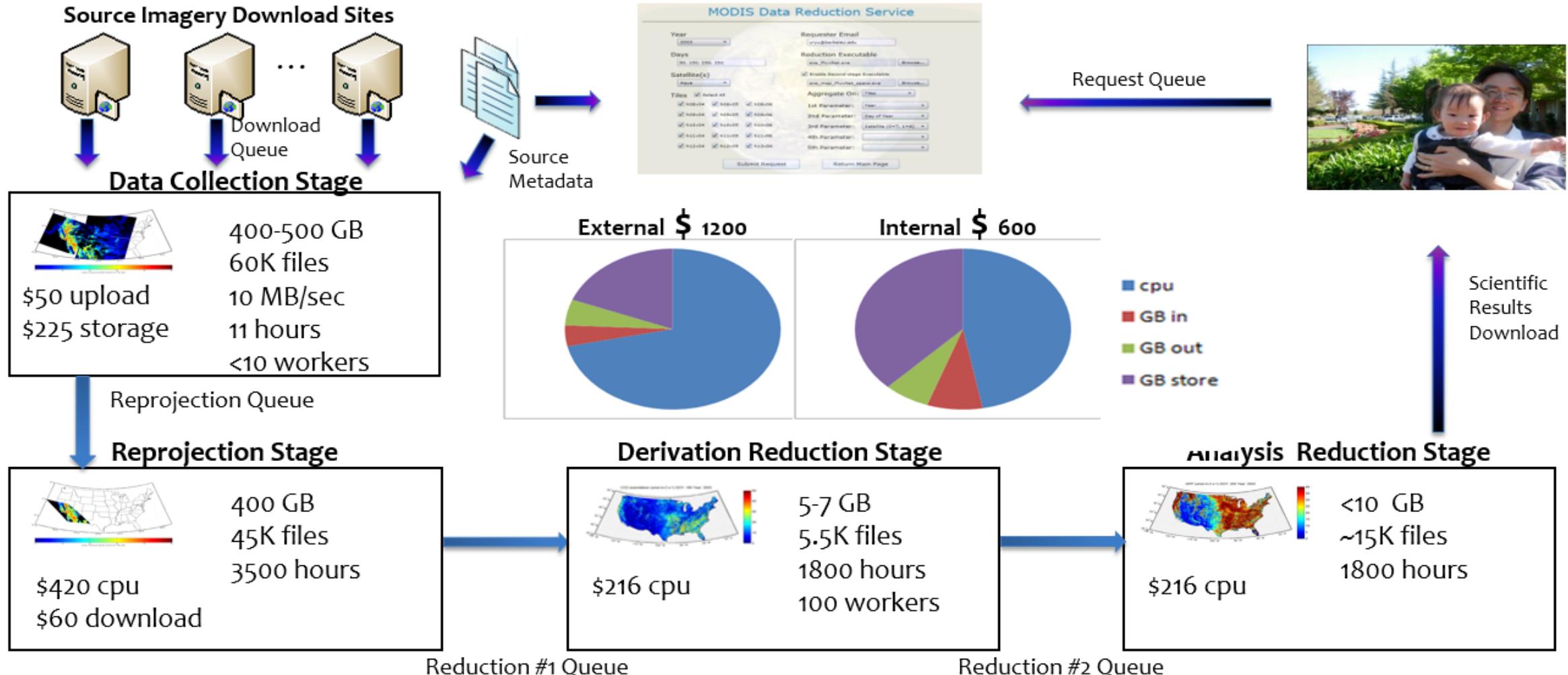


## Two MODIS satellites

- Terra, launched 12/1999
- Aqua, launched 05/2002
- Near polar orbits
- Global coverage two days
- Sensitive in 36 spectral bands

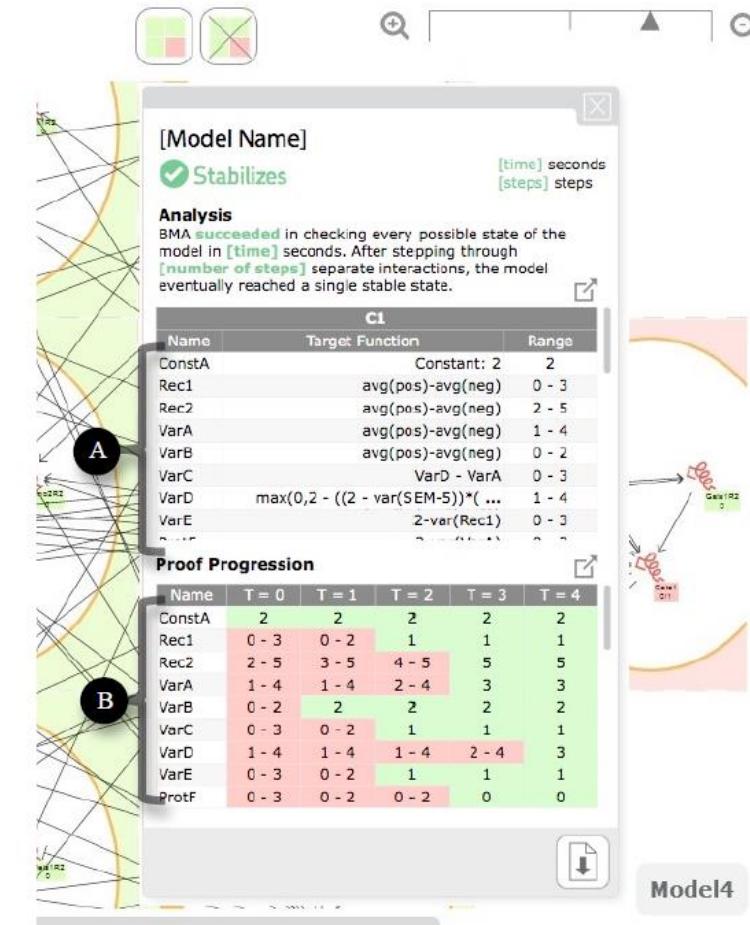
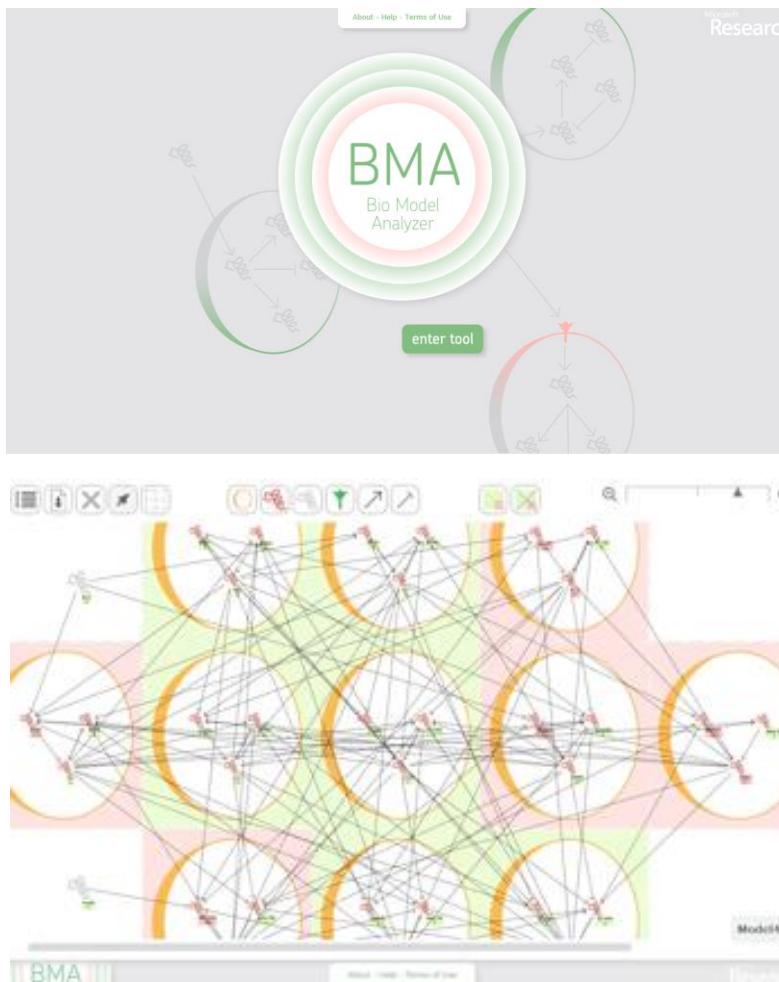


# Computing Evapotranspiration for One US Year



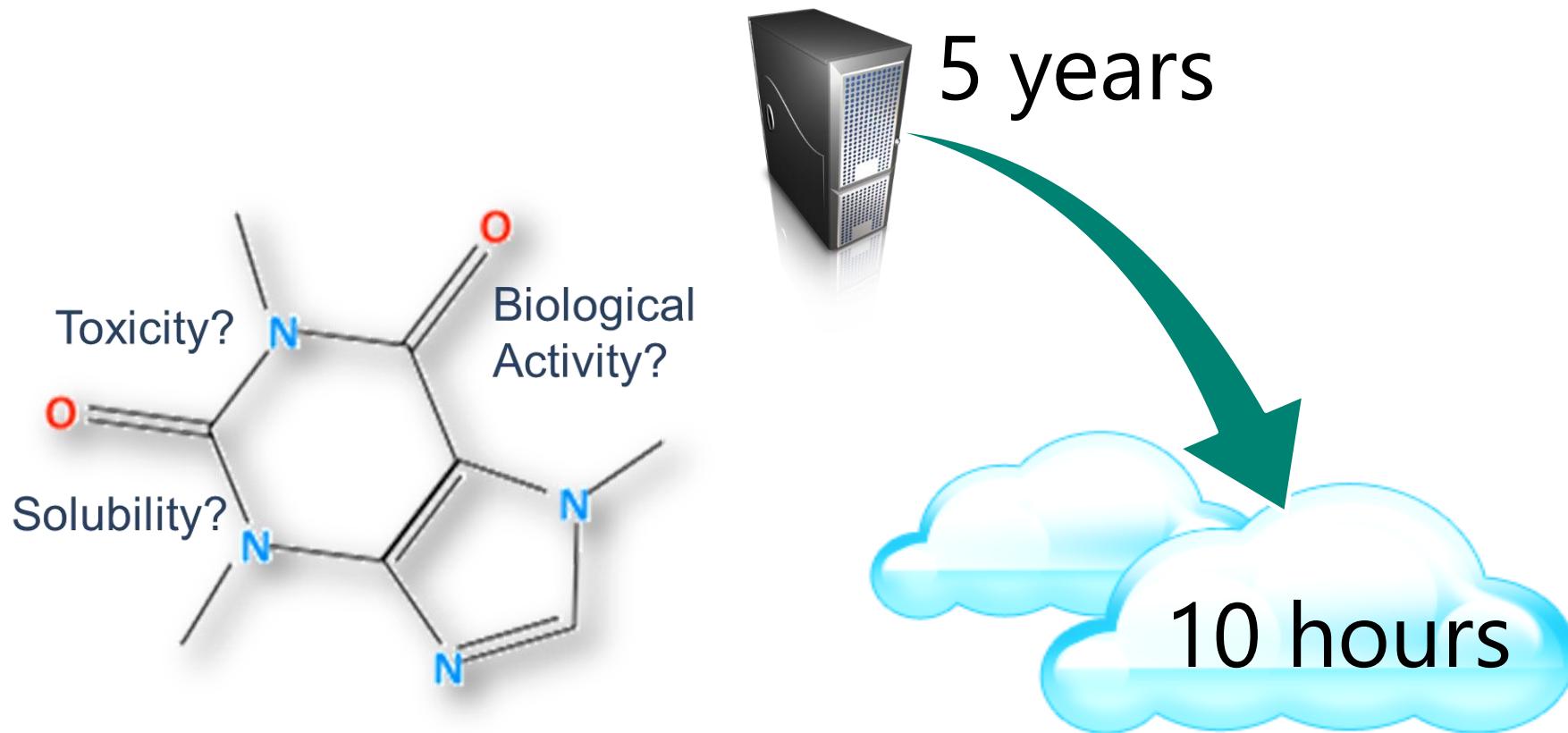
Total cost \$1,800, where all storage costs assume 3 month project duration;

# Computing Cancer



<http://biomodelanalyzer.research.microsoft.com/>  
Jasmin Fisher, Microsoft Research Cambridge

# Accelerating Drug Design - QSAR



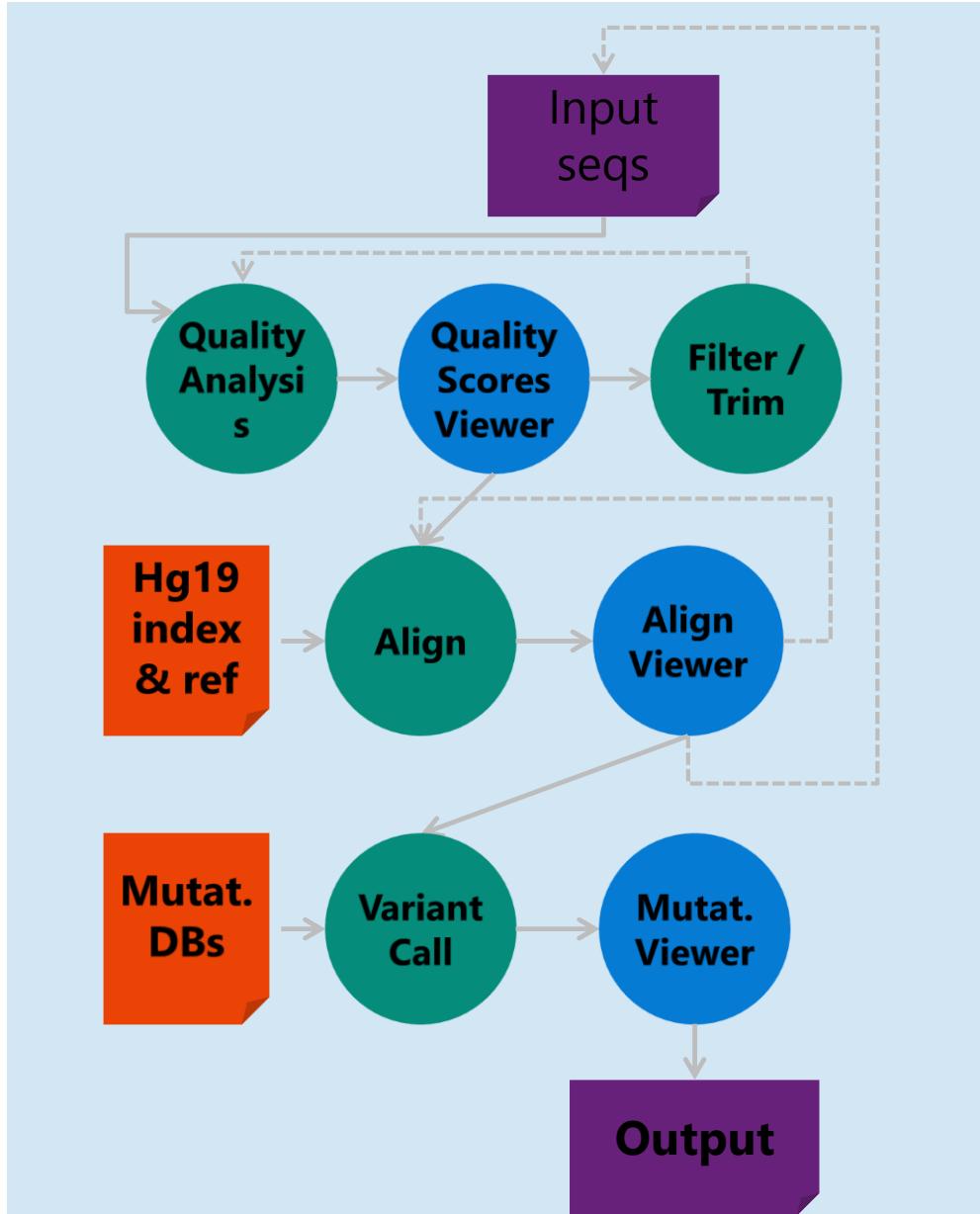
e-infrastructure



CANCER  
RESEARCH  
UK



# NGS Pipelines in the Cloud



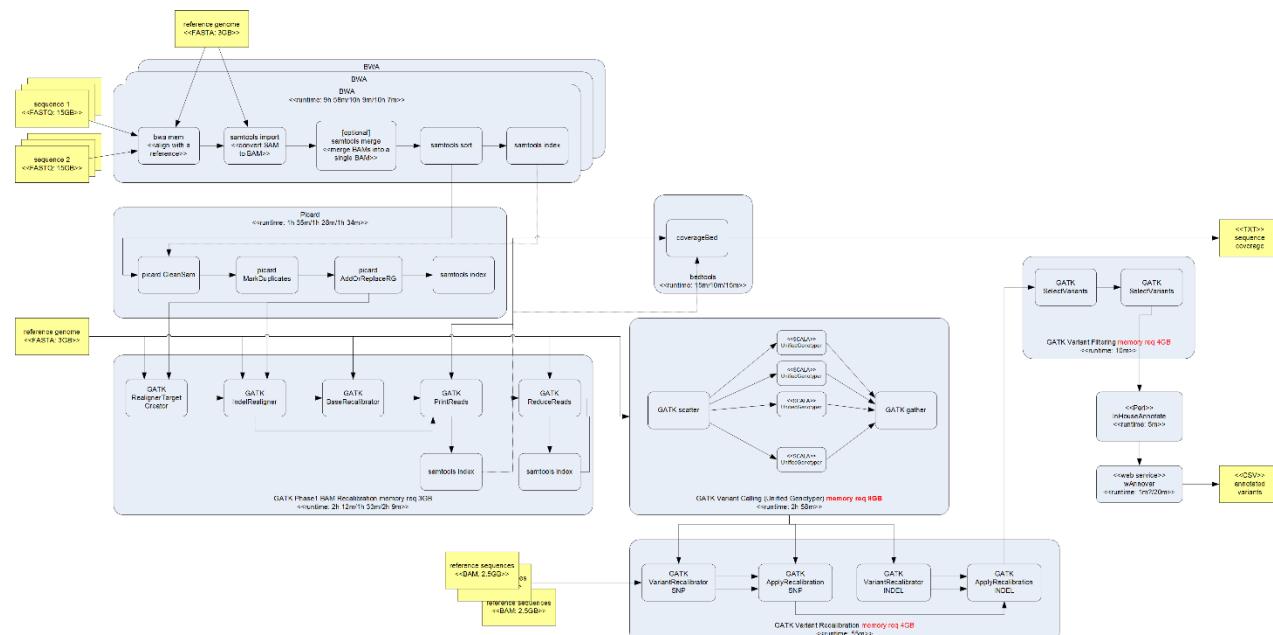
PRINCIPE FELIPE  
CENTRO DE INVESTIGACION

# Cloud e-Genome

To translate genetic testing by whole-exome sequencing into clinical practice

## Objectives:

- **Cost, Scalability:** Demonstrate the cost-effectiveness of whole-exome data processing pipelines at population scale
- **Usability:** Demonstrate a user-facing tool for variant interpretation and genetic diagnosis by clinicians



# Cloud-enabled Genomics

Sequence Quality Check results for SRR099123-1a.fastq

Wed 8 May 2013 SRR099123-1a.fastq

FastQC Report

Summary

Basic Statistics

Measure	Value
Filename	SRR099123-1a.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	50000
Filtered Sequences	0
Sequence length	36
%GC	41

Per base sequence quality

Produced by FastQC (version 0.10.1)

Filebox

- Add Local
- Del
- View

- SRR027009.fastq 86.85 MB; 4/8/2013 12:50:17 PM
- SRR099123-1a.fastq 366.07 MB; 5/8/2013 11:09:22 AM
- SRR027009.trim.fastq.sort.vcf 1.58 MB; 5/8/2013 8:08:45 AM  
Produced by e-SC [Quality Control w. Trimming]
- SRR099123-1a.fastq 8.93 MB; 5/8/2013 11:21:24 AM
- SRR099123-1.trim.fastq.sort.bam 2.03 MB; 5/8/2013 12:11:57 PM  
Produced by e-SC [Alignment]
- SRR099123-1.trim.fastq.sort.vcf 72.36 MB; 5/8/2013 12:06:55 PM  
Produced by e-SC [Alignment]
- SRR027009.trim.fastq.sort.bam 30.14 MB; 5/7/2013 2:26:58 PM  
Produced by e-SC [Alignment]
- SRR027009.trim.fastq 168.88 MB; 5/8/2013 11:50:59 AM  
Produced by e-SC [Quality Control w. Trimming]
- SRR027009.trim.fastq 99.98 MB; 5/7/2013 8:11:09 PM  
Produced by e-SC [Quality Control w. Trimming]

Sequence Quality Check action completed successfully.

Copyright © Cloud4Science project 2013

Sequence Alignment results

Cloud-4-Science Web Portal 1.0.50 | Home About

Toolbox

Check Seq. Quality Trim & Filter Align Detect Variants Generate Reference Index

Region overview

Detailed Information

Filebox

- Add Local
- Del
- View

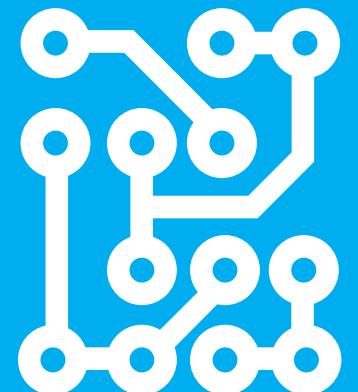
- SRR027009.fastq 86.85 MB; 4/8/2013 12:50:17 PM
- SRR099123-1a.fastq 366.07 MB; 5/8/2013 11:09:22 AM
- SRR099123-1.trim.fastq.sort.bam 2.03 MB; 5/8/2013 12:11:57 PM  
Produced by e-SC [Alignment]
- SRR099123-1.trim.fastq.sort.vcf 72.36 MB; 5/8/2013 12:06:55 PM  
Produced by e-SC [Alignment]
- SRR027009.trim.fastq.sort.bam 30.14 MB; 5/7/2013 2:26:58 PM  
Produced by e-SC [Alignment]
- SRR027009.trim.fastq 168.88 MB; 5/8/2013 11:50:59 AM  
Produced by e-SC [Quality Control w. Trimming]
- SRR027009.trim.fastq 99.98 MB; 5/7/2013 8:11:09 PM  
Produced by e-SC [Quality Control w. Trimming]

Sequence Alignment action completed successfully.

Copyright © Cloud4Science project 2013

# Linux Offering

- Linux as a first class citizen in Azure
- Open Sourcing critical components
- Documenting API
- We will offer both Community
- and Commercial Distributions
- You will be able to buy support for the commercial distributions



ALL

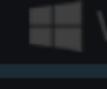
# Virtual machine operating system selection

WEB  
0

ALL

VIRT  
5MOB  
0CLOUD  
16SQL  
7STO  
26HDI  
2

NET



Wind



Wind



ALL

PLATFORM IMAGES

MY IMAGES

MY DISKS



Windows Server 2008 R2 SP1



Windows Server 2012 Datacenter



OpenLogic CentOS 6.3



openSUSE 12.3



RightScale Linux v13



SUSE Linux Enterprise Server 11 SP2



Ubuntu 12.04



Ubuntu 12.10



wenmingsaved



whitehall



boothdemo1-boothdemo1-0-2012



Microsoft SQL Server...

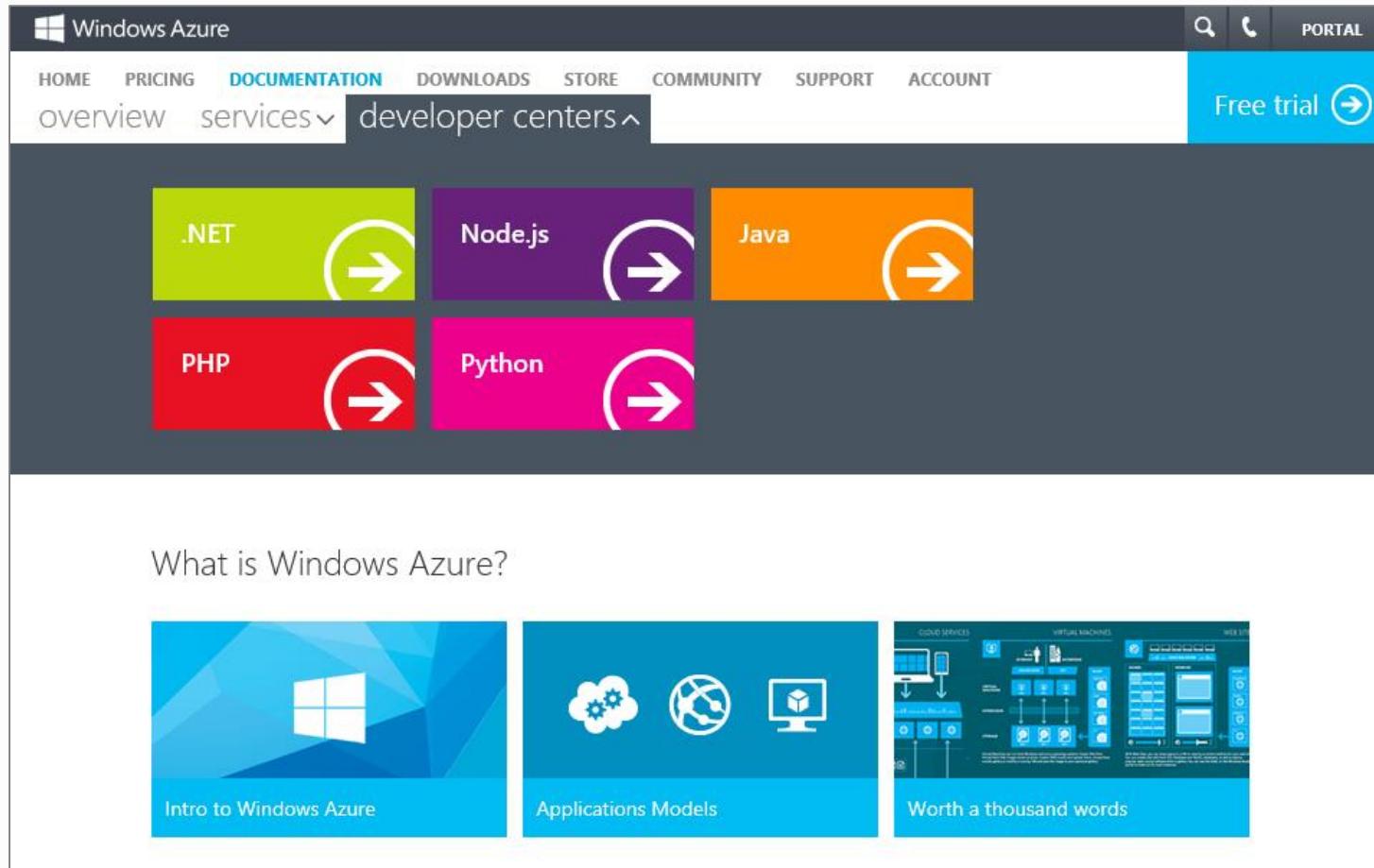
SQL Server 2012 SP1 Cumulative Update 2 Evaluation Edition (64-bit) on Windows Server 2008 R2 Service Pack 1. Virtual Machines created by using this SQL Server Evaluation Edition will expire on August 20, 2013. This image contains the full version of SQL Server. Some SQL Server 2012 components require additional setup and configuration before use. Medium is the minimum recommended virtual machine size for this image. To evaluate the advanced capabilities of SQL Server 2012, we recommend that you use a virtual machine size of Large or Extra Large.

PUBLISHER Microsoft SQL Server Group

OS FAMILY Windows

LOCATION East Asia; Southeast Asia; North Europe; West Europe; East US; West US





→ <http://WindowsAzure.com>

Microsoft Azure

# Multiple/Any languages (Fortran too)



ALL ITEMS



WEB SITES

0



VIRTUAL MACHINES

2



MOBILE SERVICES

0



CLOUD SERVICES

2



SQL DATABASES

5



STORAGE

18



HDINSIGHT

0



NETWORKS

0



SQL REPORTING

0



ADD-ONS

0

all items

NAME	TYPE	STATUS	SUBSCRIPTION	LOCATION
weatherservice	SQL Database	✓ Online	WA TPM Subscription	North Central US
blackscholes	SQL Database	✓ Online	Subscription-1	North Europe
video_server	SQL Database	✓ Online	Subscription-1	West US
blackscholes4	SQL Database	✓ Online	Subscription-1	West US
wenir1	SQL Database	✓ Online	Subscription-1	North Central US
CraigDevBox	Virtual machine	⌚ Retrieving st...	WA TPM Subscription	West US
dpedemo2	Virtual machine	⌚ Retrieving st...	WA TPM Subscription	West US
serverperf	Cloud service	✓ Created	WA TPM Subscription	West US
weatherservice	Cloud service	⌚ Retrieving S...	WA TPM Subscription	North Central US
ckittervms	Storage Account	✓ Online	WA TPM Subscription	West US
hadoopstore	Storage Account	✓ Online	WA TPM Subscription	West US
hadoopstore1	Storage Account	✓ Online	WA TPM Subscription	West US
portalvhds1st00d071vgg3	Storage Account	✓ Online	WA TPM Subscription	West US
portalvhdsvg4n2tsscqpgm	Storage Account	✓ Online	WA TPM Subscription	North Europe



EXPORT



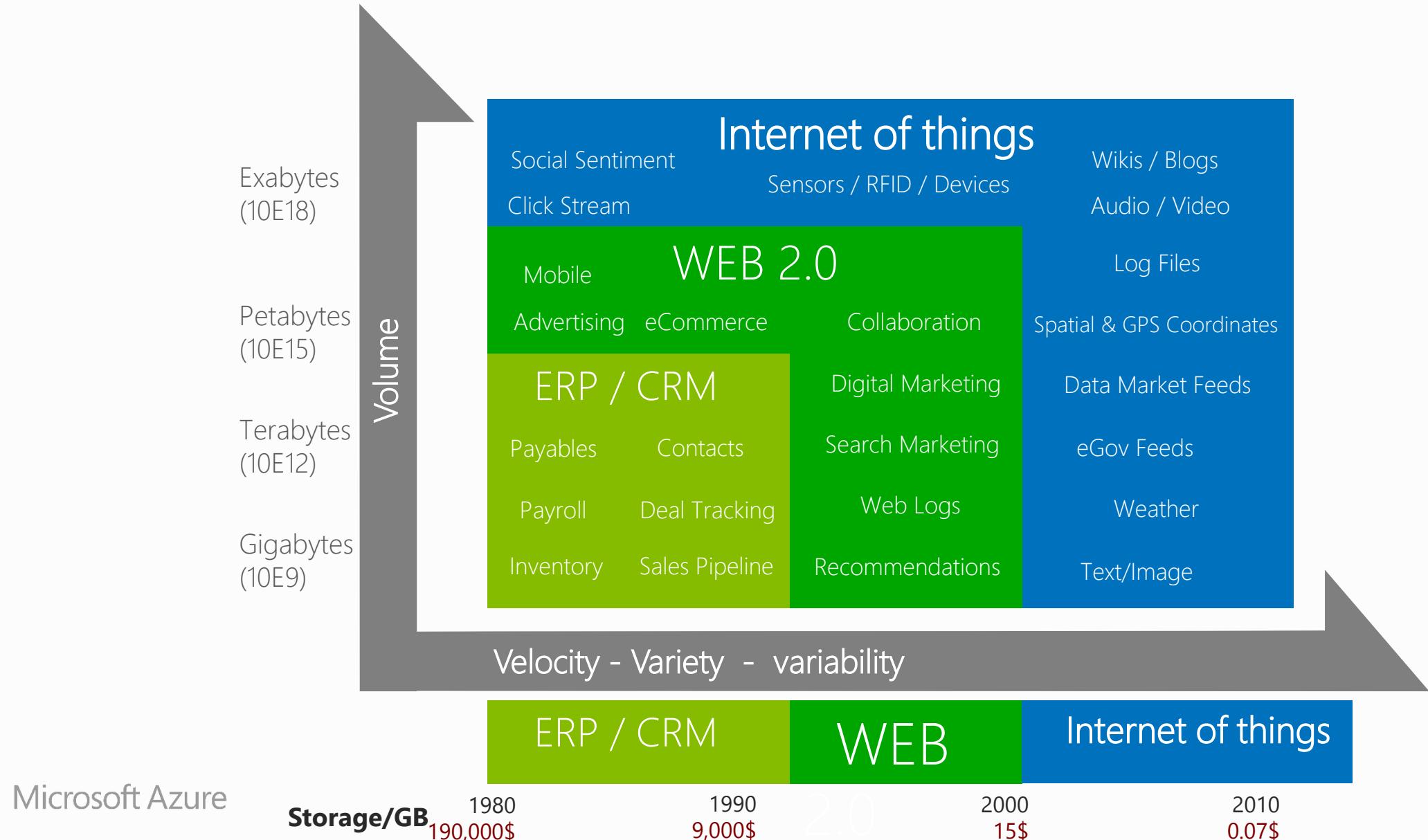
MANAGE



DELETE



# Big Data



# Predictive Analytics

Predicting future performance from historical data

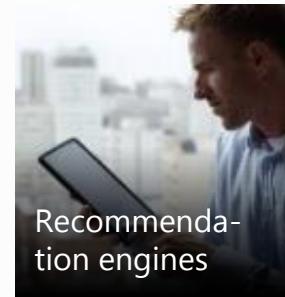
Predictive analytics  
should address the  
likelihood of something  
happening in the future,  
even if it is just an  
instant later\*



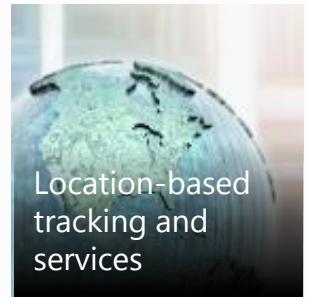
Churn analysis



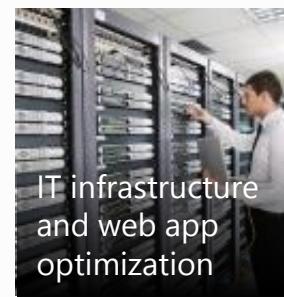
Social network analysis



Recommendation engines



Location-based tracking and services



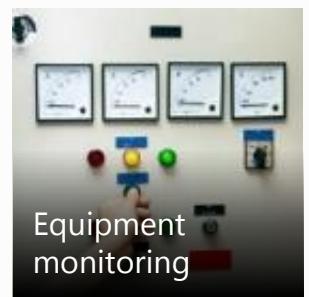
IT infrastructure and web app optimization



Weather forecasting for business planning



Legal discovery and document archiving



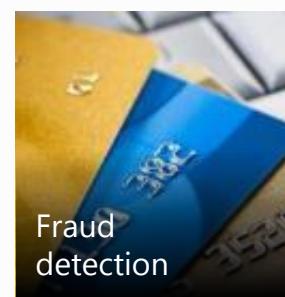
Equipment monitoring



Advertising analysis



Pricing analysis



Fraud detection



Personalized Insurance

**“**Azure Machine Learning offers a data science experience that is directly accessible to business analysts and domain experts, reducing complexity and broadening participation through better tooling.**”**

**Hans Kristiansen**

Capgemini

## Azure Machine Learning How it works

Enable custom predictive analytics solutions at the speed of the market

# One solution for Machine Learning — from data to results



**Business users easily access results:  
from anywhere, on any device**

## **ML API service** and the Developer

- Tested models available as an url that can be called from any end point

## **Azure Portal & ML API service**

and the Azure Ops Team

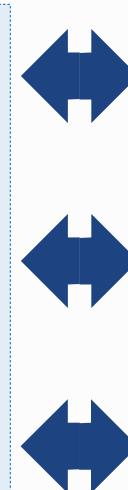
- Create ML Studio workspace
- Assign storage account(s)
- Monitor ML consumption
- See alerts when model is ready
- Deploy models to web service



## **ML Studio**

and the Data Scientist

- Access and prepare data
- Create, test and train models
- Collaborate
- One click to stage for production via the API service



HDInsight

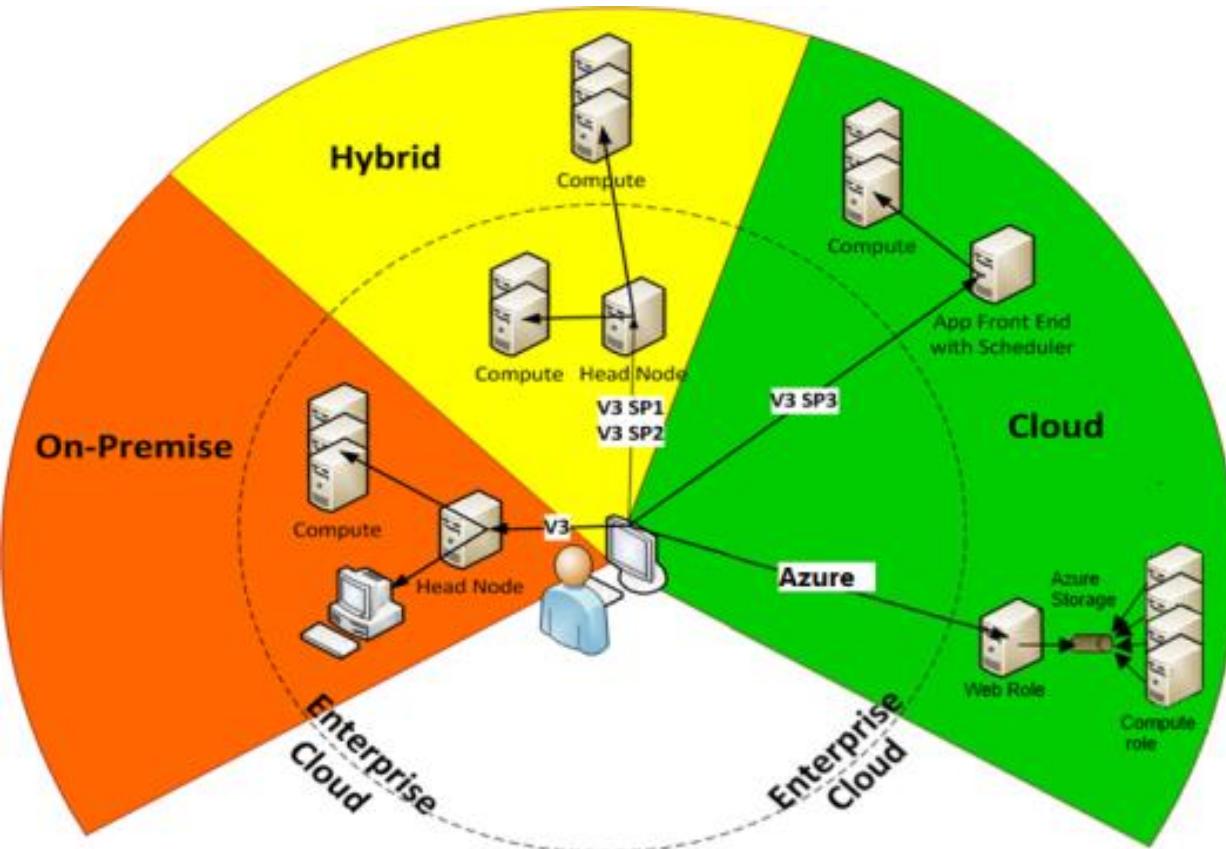


Azure Storage



Desktop Data

# Microsoft HPC Scenarios



## On-Premises

Your own servers inside your enterprise cloud.

## Hybrid

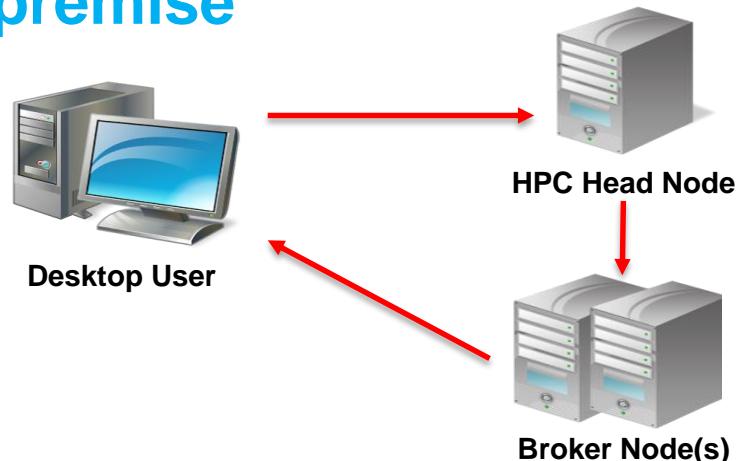
Burst from enterprise to cloud.

## Cloud

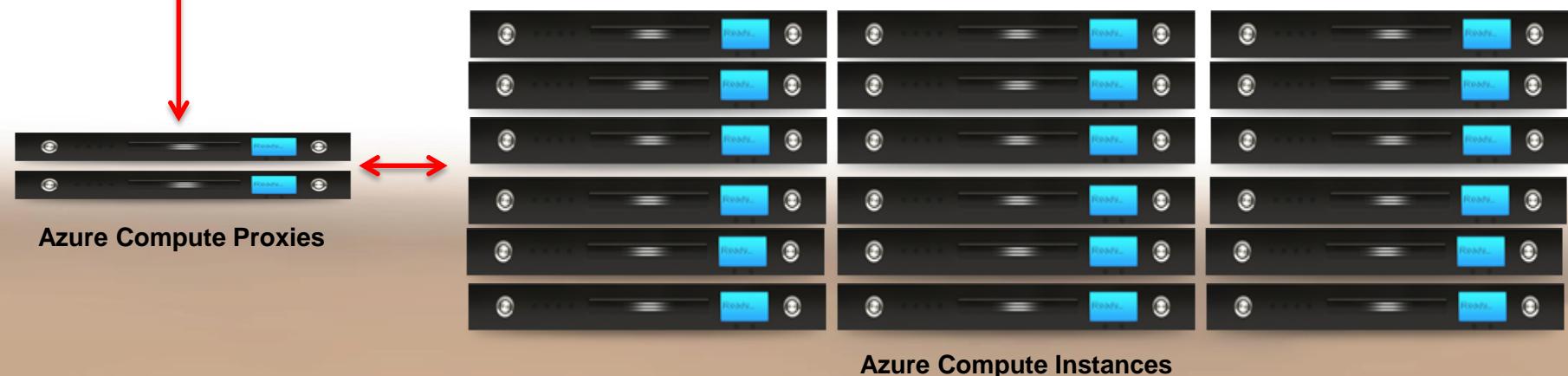
Head node in cloud or compute in cloud  
Or both.

# HPC Components (On Premises View)

## On-premise

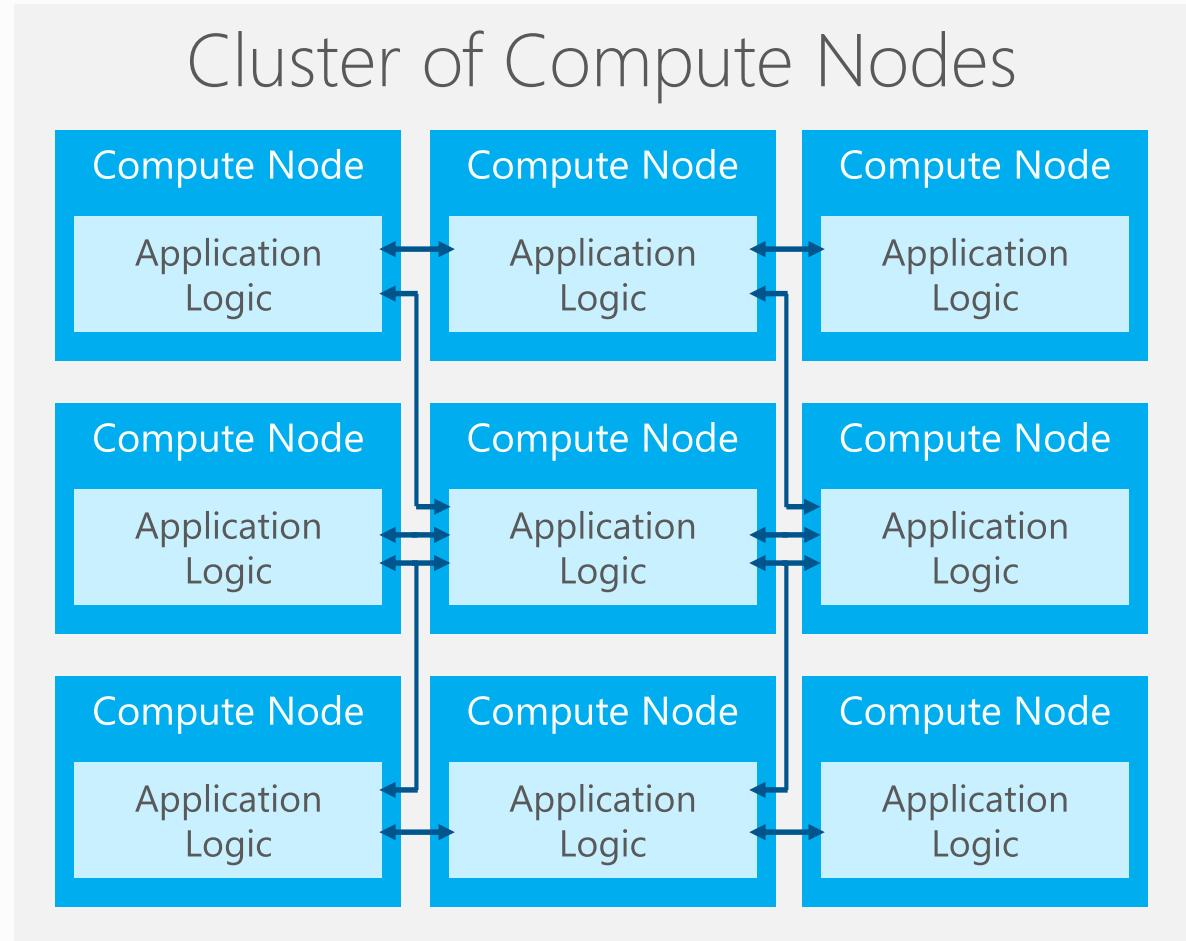
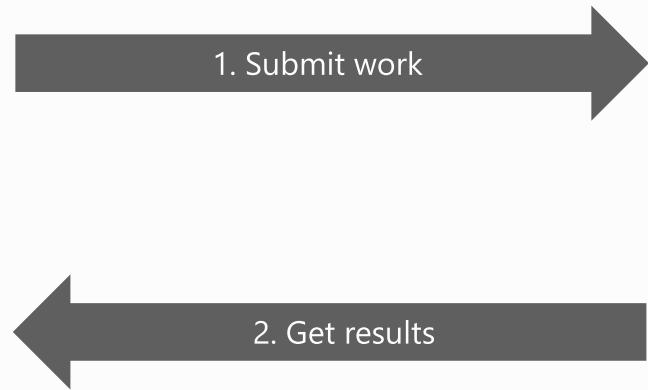


## Microsoft Azure



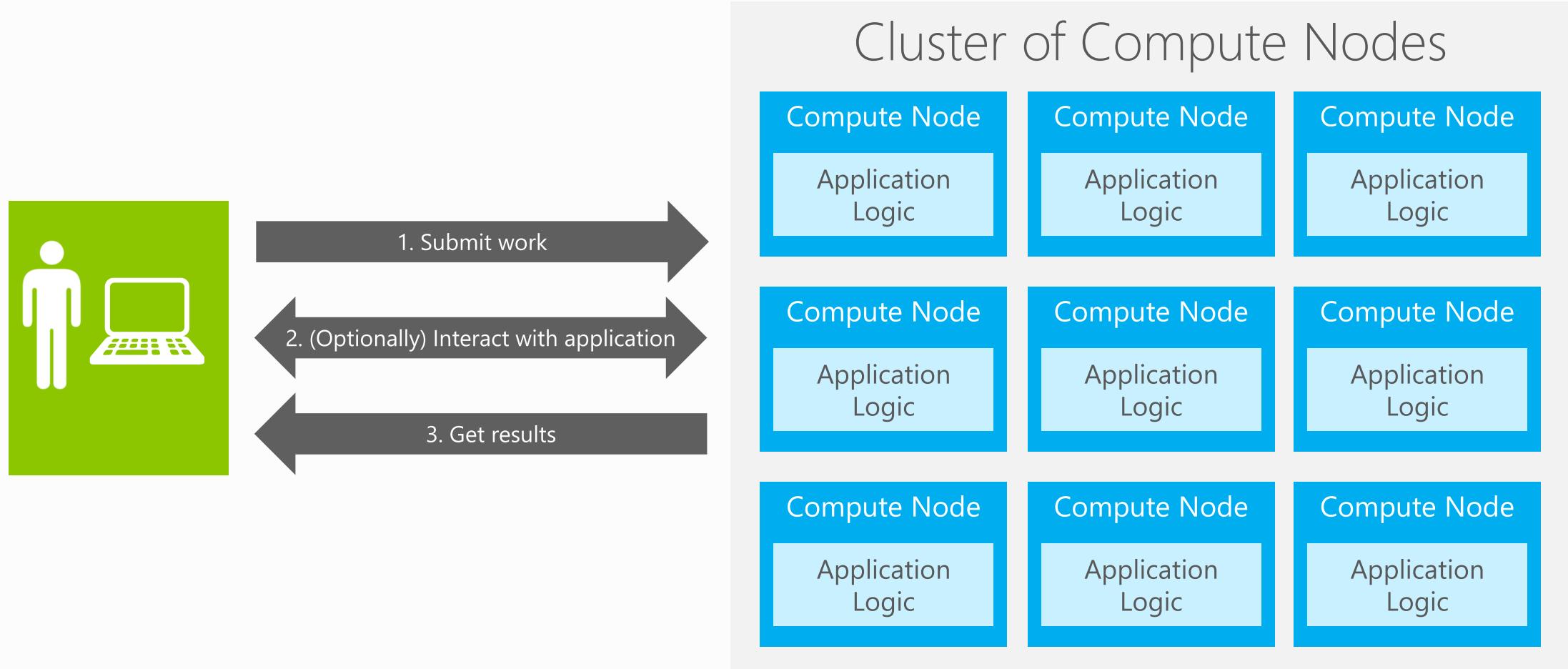
# Tightly Coupled App (MPI)

An illustration



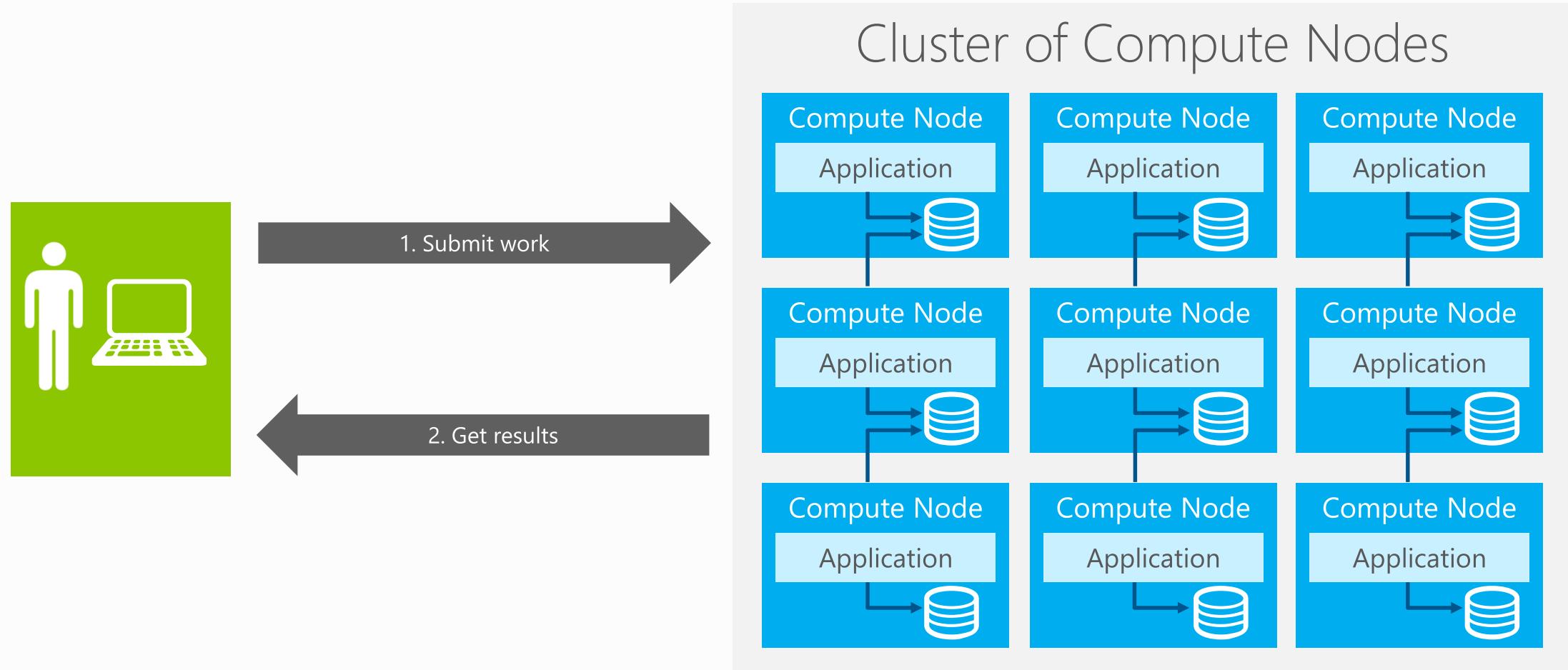
# Embarrassingly Parallel Apps

An illustration



# IO Intensive Applications (HDInsight)

An illustration



# HPC Software Workload Types



Message Passing Interface

Low level API and executable launch  
Low level network access uses NetworkDirect  
Creates a cluster by assignment of communicators



Parametric Sweep

Command Line Interface  
Can be written in any language hostable in Windows  
Can be scheduled in a robust way

The choice of HPC Software type depends on the workload of the expected tasks for the Cluster. Each has its areas of strength, from development simplicity to pure performance power.

## ■ Additional slides

# Microsoft Hadoop Vision

Insights to all users by activating new types of data



# MICROSOFT HADOOP STRATEGY

Make Hadoop Enterprise Ready



Apache distribution of Hadoop,  
partner w/ HortonWorks

Submit changes back to Apache Foundation

Optimized for Windows  
& Azure

'Just works' on Windows Azure and Server

Wider Ecosystem

Integration with Visual Studio, Javascript, Excel, etc.

Enterprise Readiness

Performance, Scale, High Availability  
Management, Ease of use  
Security, Data Governance  
Integration with Active Directory and System Center

Structured and Unstructured

Integrate as part of our overall data platform

# HADOOP ON ...

## Hadoop On Windows

Dedicated workload cluster  
HDFS is the persistent data store

## Hadoop on Azure

Transient clusters, keep as long as you want  
Persistent data in Azure storage  
HDFS is a cache

## Hadoop On Windows (Virtualized)

Cluster nodes on Hyper-V  
Multi-workload cluster  
Persistent data to network attached storage

# Introducing the Microsoft Analytics Platform System

Your turnkey modern data warehouse appliance

SQL Server PDW appliance is now called: Microsoft Analytics Platform System (APS)

Besides SQL Server PDW a new workload can be run in the same appliance: HDInsight (Hadoop) - optionally

Configuration 1: SQL Server PDW only

Configuration 2: a minimum of 2 (= 1x HP Scale Unit) or 3 (= 1x Dell Scale Unit) SQL Server PDW nodes + additional HDInsight Scale Unit(s)

Enabled with SQL Server 2012 PDW Appliance Update 1 (AU1) including additional PDW specific features

# Introducing the Microsoft Analytics Platform System

Your turnkey modern data warehouse appliance

## Enterprise-ready big data

- PDW and HDInsight in a single appliance
- Enterprise-ready Hadoop
- Integrated querying across Hadoop and PDW using T-SQL
- Big data insights to a billion users

## Next-generation performance at scale

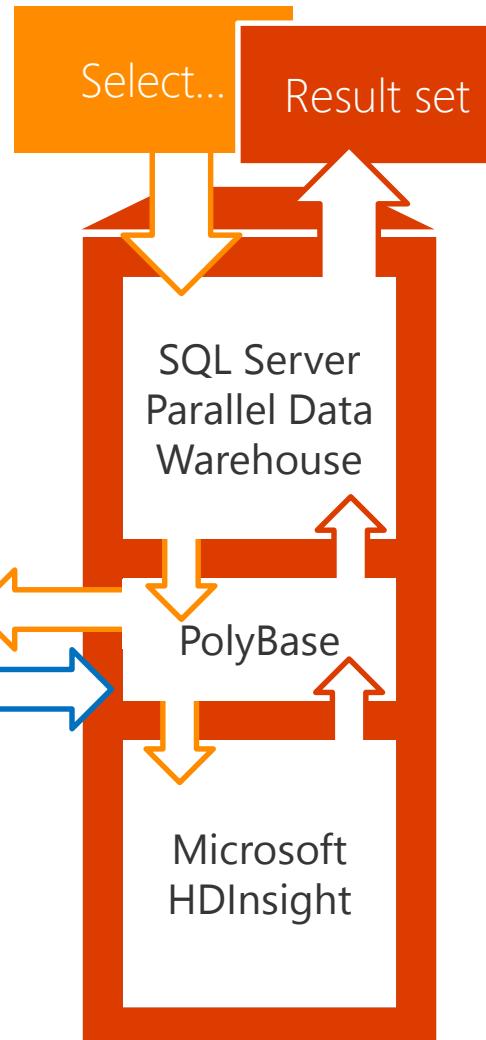
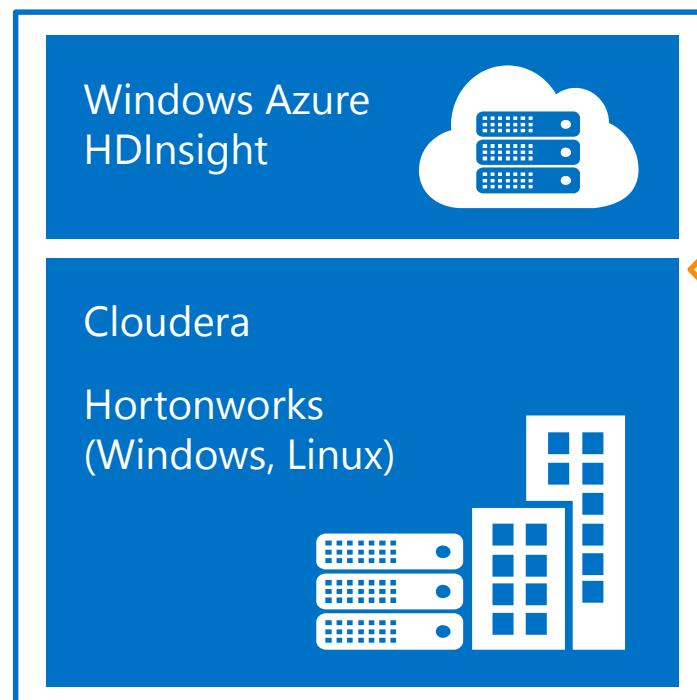
- Near real-time performance with In-Memory
- Scale-out to accommodate your growing data
- Remove DW bottlenecks with MPP SQL Server
- Concurrency that fuels rapid adoption

## Engineered for optimal value

- APS provides the industry's lowest DW price/TB
- Value through a single appliance solution
- Value with flexible hardware options using commodity hardware

# Query Hadoop data with T-SQL using PolyBase

Bringing the worlds of big data and the data warehouse together for users and IT



Single T-SQL query model for PDW and Hadoop with rich features of T-SQL including joins without ETL

Leverages the power of MPP to enhance query execution performance

Supports Windows Azure HDInsight to enable new hybrid cloud scenarios

Query non-Microsoft Hadoop distributions such as Hortonworks and Cloudera

# Native Query Across Hadoop and PDW

## Polybase Features in SQL Server PDW

- Querying data in Hadoop from PDW using regular **SQL queries**, including
  - Full SQL query access to data stored in HDFS, represented as 'external tables' in PDW
    - Basic statistics support for data coming from HDFS
  - Querying across PDW and Hadoop tables (joining 'on the fly')
- Fully parallelized, **high performance import** of data from HDFS files into PDW tables
- Fully parallelized, **high performance export** of data in PDW tables into HDFS files
- Integration with various Hadoop distributions: Hadoop on Windows Server, Hortonwork and Cloudera.
- Supporting Hadoop 1.0 and 2.0

# External Tables (Polybase tables)

- » An external table is PDW's representation of data residing in HDFS
- » The “table” (metadata) lives in the context of a SQL Server database
- » The actual table data resides in HDFS

```
CREATE EXTERNAL TABLE table_name ({<column_definition>} [,...n ])  
    {WITH (LOCATION = '<URI>', [FORMAT_OPTIONS = (<VALUES>) ]) }  
[;]
```

Required to indicate  
location of Hadoop cluster

Optional format options  
associated with parsing of data  
from HDFS (e.g. field delimiters  
& reject-related thresholds)

# PDW Hadoop use cases & examples

## [1] Retrieve data from HDFS with a PDW query

- » Seamlessly join structured and semi-structured data

```
SELECT Username FROM ClickStream c, User u WHERE c.UserID = u.ID  
AND c.URL='www.bing.com';
```

## [2] Import data from HDFS to PDW

- » Parallelized CREATE TABLE AS SELECT (CTAS)
- » External tables as the source
- » PDW table, either replicated or distributed, as destination

```
CREATE TABLE ClickStreamInPDW WITH DISTRIBUTION = HASH(URL)  
AS SELECT URL, EventDate, UserID FROM ClickStream;
```

## [3] Export data from PDW to HDFS

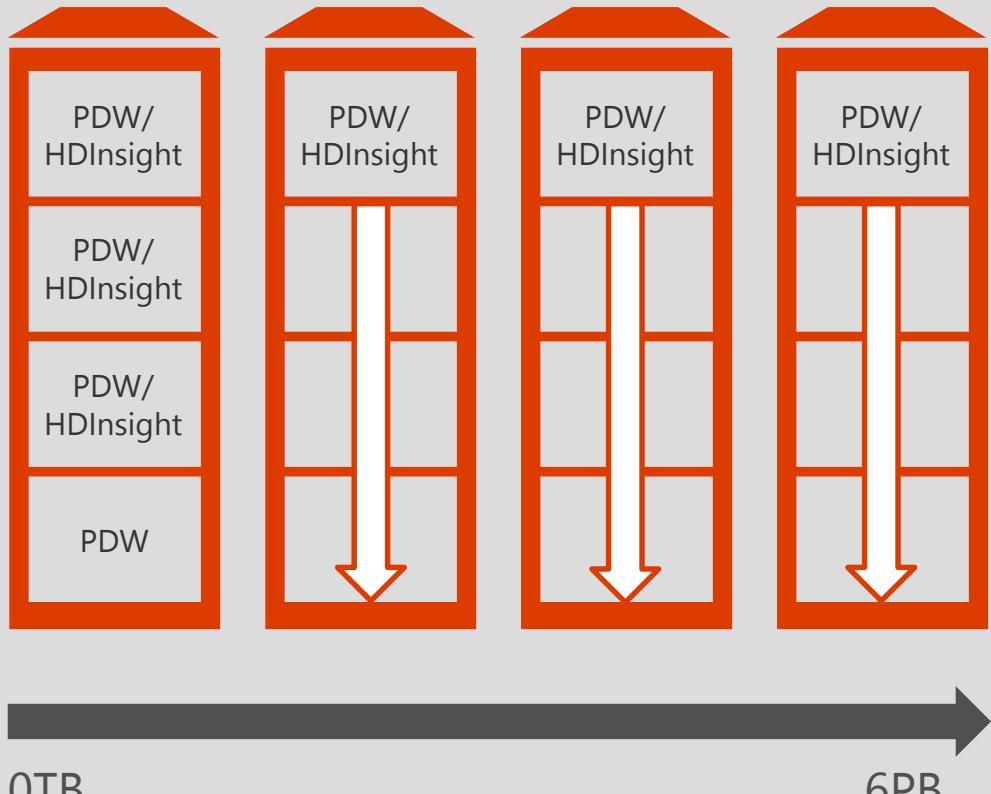
- » Parallelized CREATE EXTERNAL TABLE AS SELECT (CETAS)
- » External table as the destination; creates a set of HDFS files

```
CREATE EXTERNAL TABLE ClickStream2 (URL, EventDate, UserID)  
WITH (LOCATION ='hdfs://MyHadoop:5000/joe', FORMAT_OPTIONS (...))  
AS SELECT URL, EventDate, UserID FROM ClickStreamInPDW;
```

# Scaling out relational data to petabytes

## Scale-out technologies in the Analytics Platform System

### Scale-out



Multiple nodes with dedicated CPU, memory, and storage

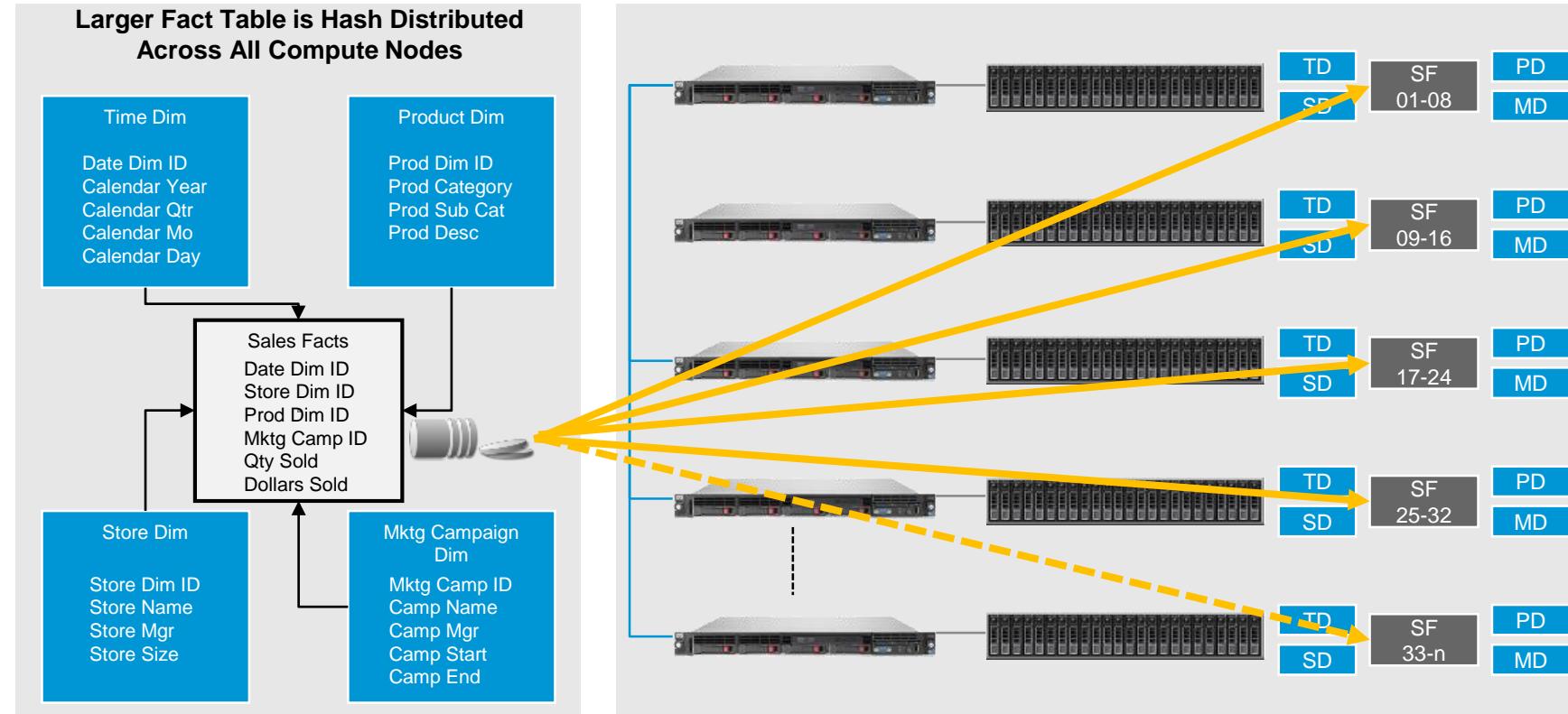
Ability to incrementally add hardware for near-linear scale to multiple petabytes

Ability to handle query complexity and concurrency at scale

No “forklift” of prior warehouse to increase capacity

Ability to scale out HDInsight and PDW

# Ultra Shared Nothing architecture: Distribution



## HP Base Rack

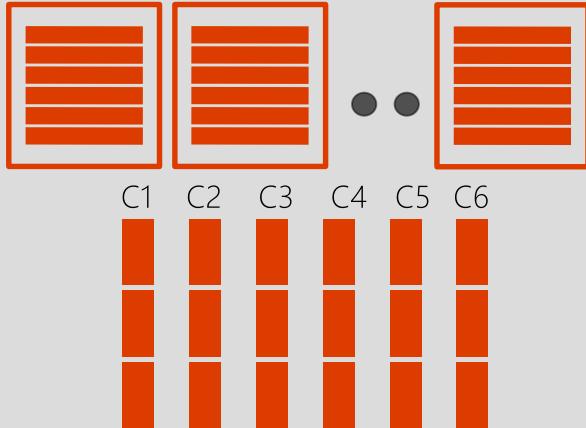
The following diagram shows the architecture for an HP Base rack. This Base rack ships with 2 Compute nodes which have the same architecture as a Scale unit. To add more Compute nodes to the Base rack, you can add Scale units. A full HP Base rack has 8 Compute nodes.



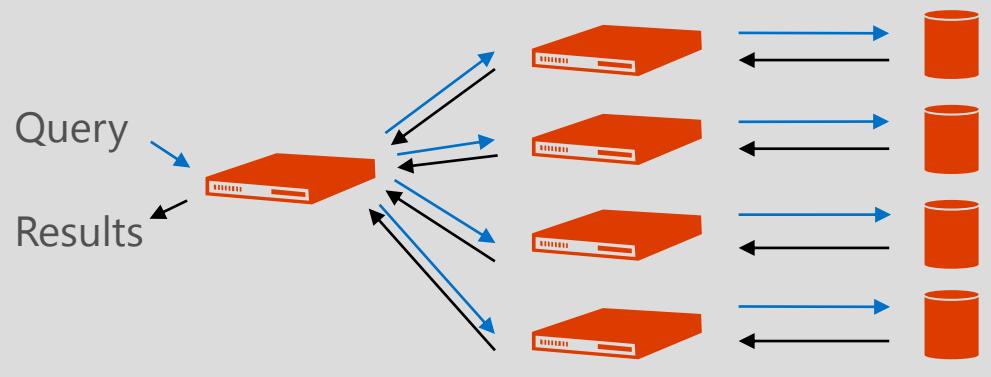
# Blazing fast performance

MPP and In-memory columnstore for next-generation performance

## Columnstore index representation



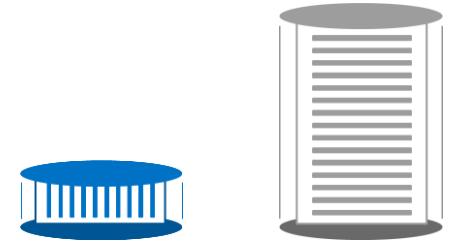
## Parallel query execution



Up to 100x  
faster queries



Up to 15x  
more compression

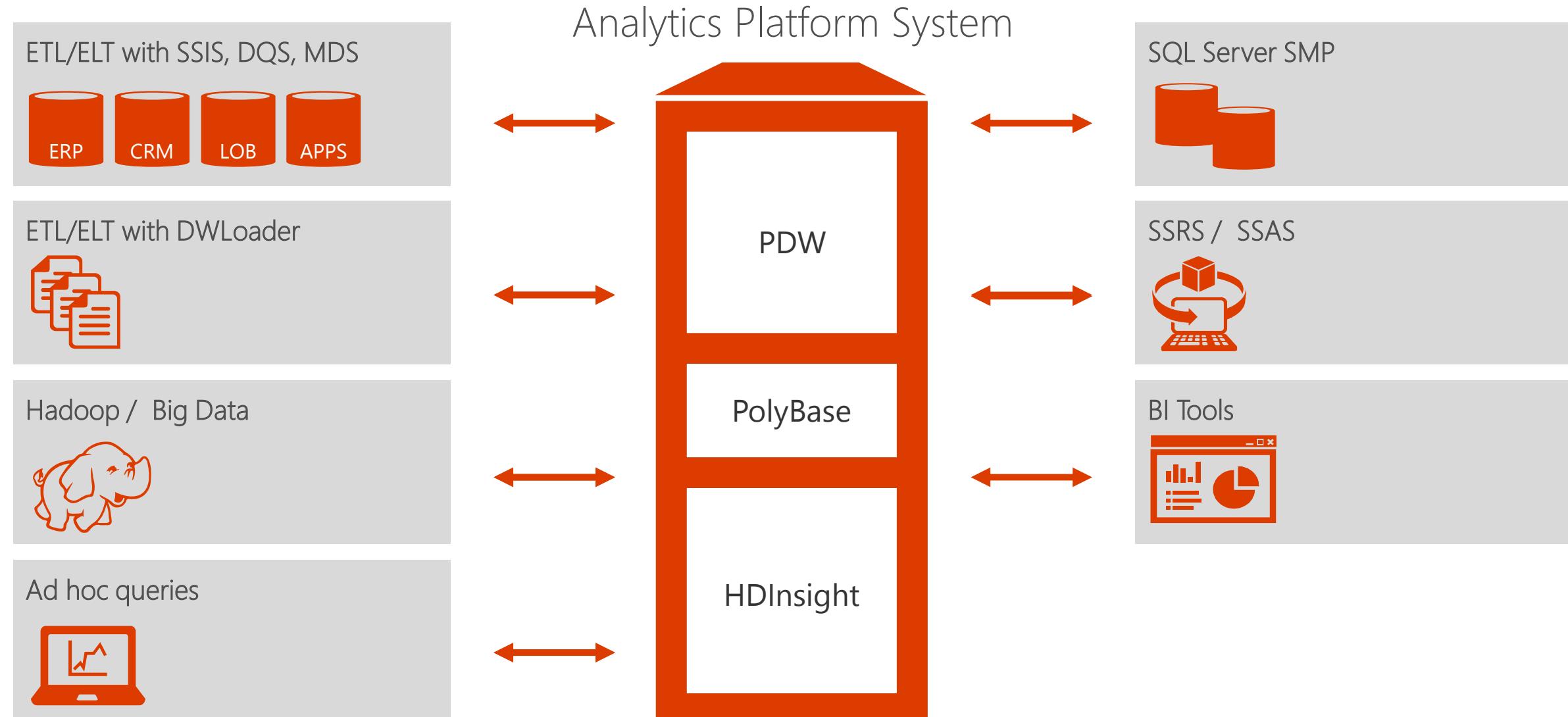
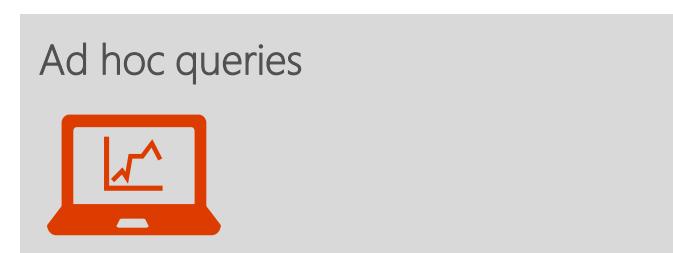
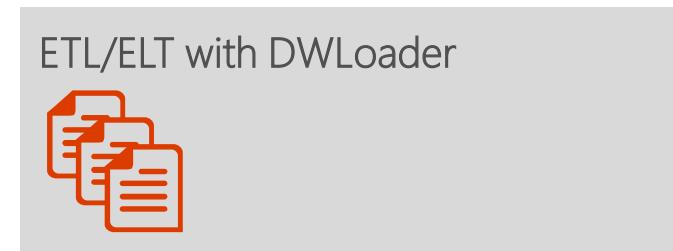


Updatable clustered columnstore vs. table with customary indexing

- Store data in columnar format for massive compression
- Load data into or out of memory for next-generation performance
- Updateable and clustered for real-time trickle loading

# Concurrency that fuels rapid adoption

Great performance with mixed workloads



# Power Query

Discover, combine, and refine Big Data, small data, and any data with Data Explorer for Excel.

- Excel add-in to enhance self-service BI
- Identify and import external data:
  - Relational dB
  - Excel
  - Text
  - XML
  - Odata
  - Web pages
  - Hadoop HDFS
- Discover relevant data by using search
- Combine and transform multiple data sources

CustomerID	CompanyName	ContactName	ContactTitle	Address
1	ALFKI	Maria Anders	Sales Representative	Obere Str. 57
2	ANATR	Ana Trujillo Emparedados y helados	Owner	Avda. de la Constitución 2222
3	ANTON	Antonio Moreno Taquería	Owner	Mataderos 2312
4	AROUT	Around the Horn	Sales Representative	120 Hanover Sq.
5	BERGS	Berglunds snabbköp	Order Administrator	Berguvsvägen 8
6	BLAUS	Blauer See Delikatessen	Sales Representative	Forsterstr. 57
7	BLONP	Blondesdöds före et fils	Marketing Manager	24, place Kleber
8	BOLID	Bólido Comidas preparadas	Owner	C/ Aragón, 67
9	BONAP	Bon app'	Owner	12, rue des Bouchers
10	BOTTM	Bottom-Dollar Markets	Accounting Manager	23 Tsawassen Blvd.
11	BSBEV	B's Beverages	Sales Representative	Fauntleroy Circus
12	CACTU	Cactus Comidas para llevar	Sales Agent	Cerrito 333
13	CENTC	Centro comercial Moctezuma	Marketing Manager	Sierras de Granada 9993
14	CHOPS	Chop-suey Chinese	Owner	Hauptstr. 29
15	COMMI	Comércio Mineiro	Sales Associate	Av. dos Lusíadas, 23
16	CONSH	Consolidated Holdings	Sales Representative	Berkeley Gardens 12 Brewery
17	DRACD	Drachenblut Delikatessen	Order Administrator	Waisenweg 21
18	DUMON	Du monde entier	Owner	67, rue des Cinquante Otages
19	EASTC	Eastern Connection	Sales Agent	35 King George
20	ERNSH	Ernst Handel	Sales Manager	Kirchgasse 6
21	FAMILIA	Familia Arquibaldo	Marketing Assistant	Rua Orós, 92
22	FISSA	FISSA Fabrica Inter. Salchichas S.A.	Accounting Manager	C/ Moralzarzal, 86
23	FOLIG	Folies gourmandes	Assistant Sales Agent	184, chaussée de Tournai
24	FOLKO	Folk och fä HB	Owner	Åkergratan 24
25	FRANK	Frankenversand	Marketing Manager	Berliner Platz 43
26	FRANR	France restauration	Marketing Manager	54, rue Royale
27	FRANS	Franchi S.p.A.	Sales Representative	Via Monte Bianco 34
28	FURIB	Furia Bacalhau e Frutos do Mar	Sales Manager	Jardim das rosas n. 32
29	GALED	Galería del gastrónomo	Marketing Manager	Rambla de Cataluña, 23

# Powerful Self-Service BI with Excel 2013

## Discover & Combine



Search and find internal & external data

Clean, transform, and shape data

Merge and combine data from multiple sources

## Analyze & Model



Lightning fast analytics with in-memory technology

Model relationships, custom measures, hierarchies, and KPI's

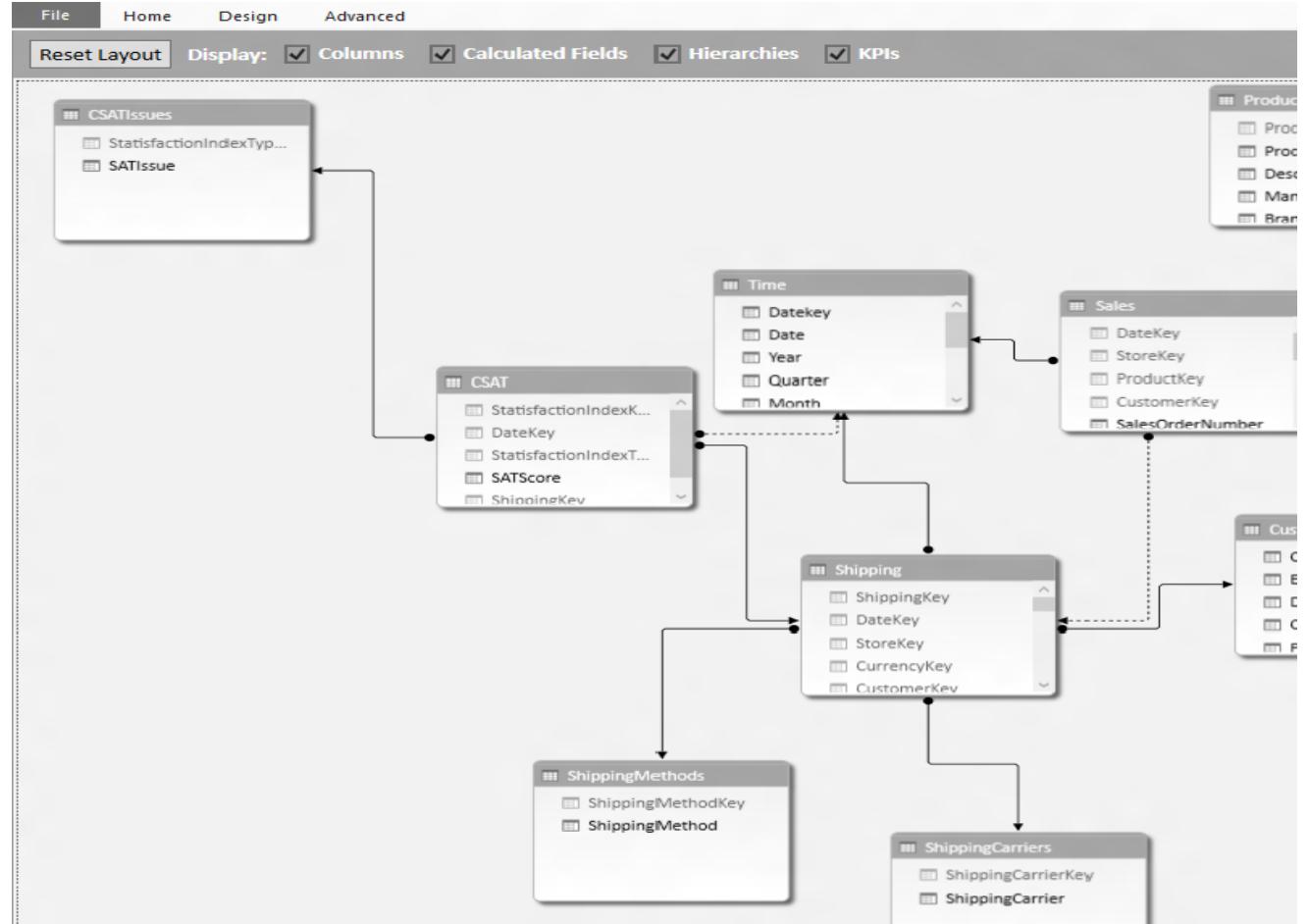
## Visualize & Explore



Bring your data to life with interactive visualization

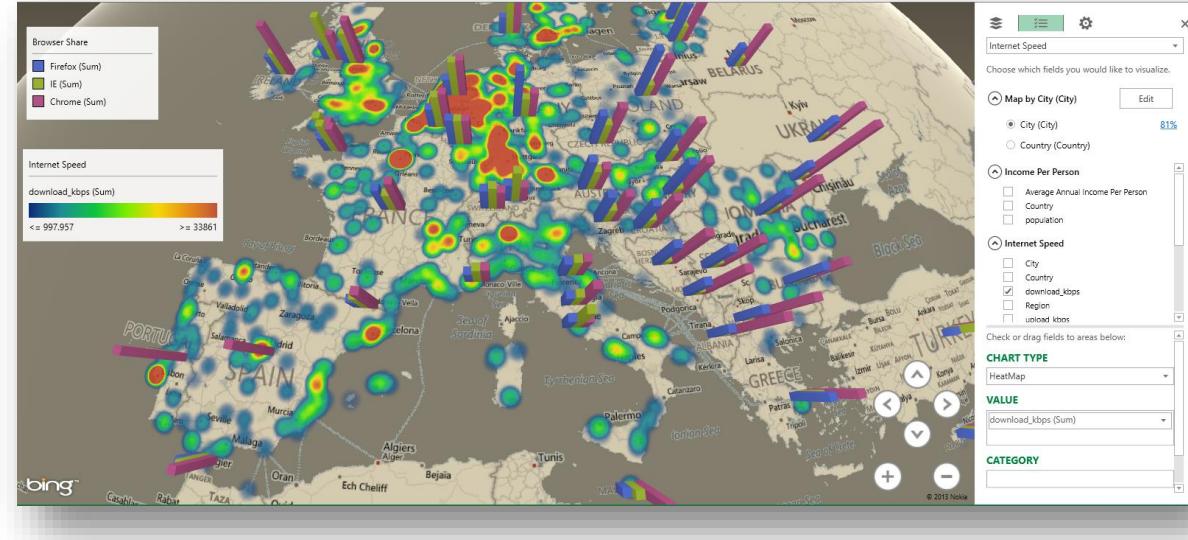
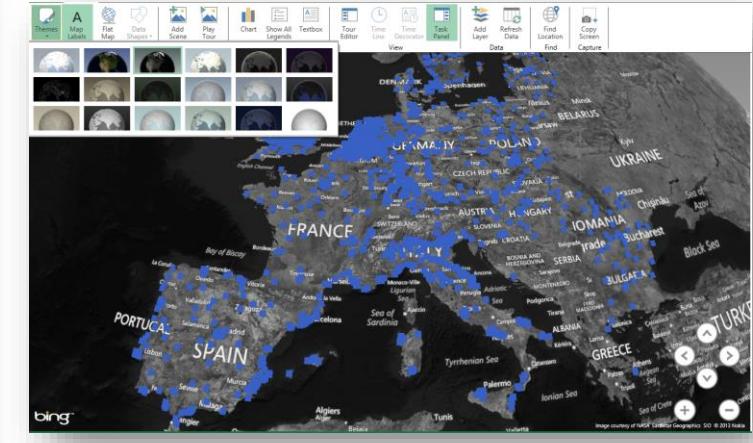
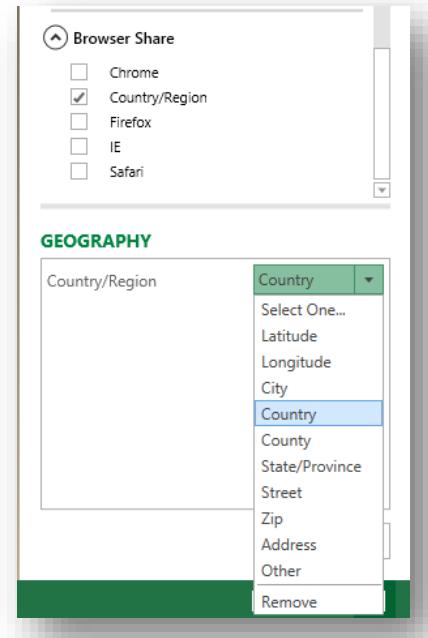
Explore data in new ways to discover hidden insights

## PowerPivot

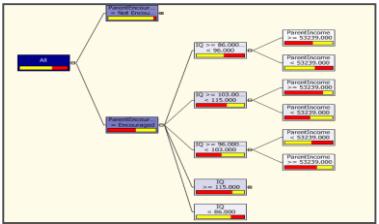


# Map Data

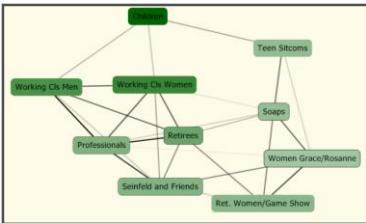
- Geospatial and temporal data in Excel
  - Calculated Fields and Hidden Columns
- Geo-code and Themed maps with Bing
- Visualize multiple layers at once
- 3D Columns, Bubble/Pies, Heat Maps
  - Regions for geopolitical boundaries



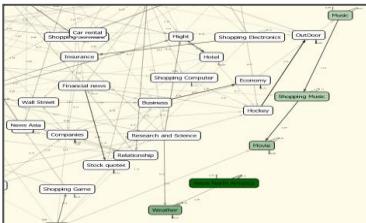
# SQL Server Data Mining Algorithms



- Decision Trees
  - The most popular data mining technique
  - Used for classification



- Clustering
  - Finds natural groupings inside data

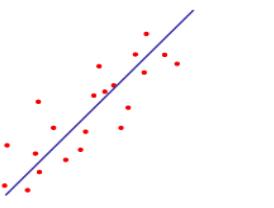


- Sequence Clustering
  - Groups a sequence of discrete events into natural groups based on similarity
  - Use this algorithm to understand how visitors use your Web site

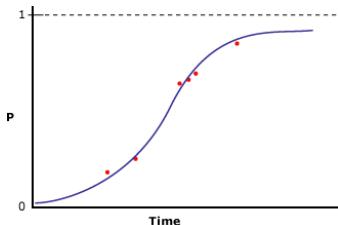
# SQL Server Data Mining Algorithms

Discrimination scores for Professional/Technical and Service Workers		
Attributes	Values	Favor: Professional/Techn., Favor: Service Workers
Education Years	15-20	██████████
Education Years	12-13	██████
Education Years	7-12	███
nelson_his/young AND THE RES..	Missing	██
nelson_his/young AND THE RES..	Existing	██
nelson_his/AS THE WORLD TURN..	Existing	██
nelson_his/AS THE WORLD TURN..	Missing	██

- Naïve Bayes
  - Used for classification in similar scenarios to Decision Trees

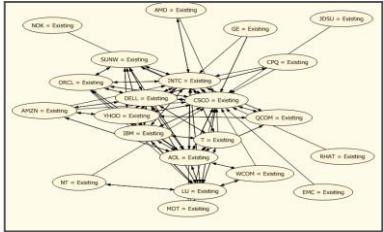


- Linear Regression
  - Finds the best possible straight line through a series of points
  - Used for prediction analysis

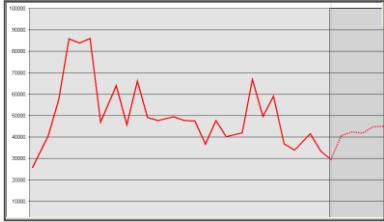


- Logistic Regression
  - Fits to an exponential factor
  - Used for prediction analysis

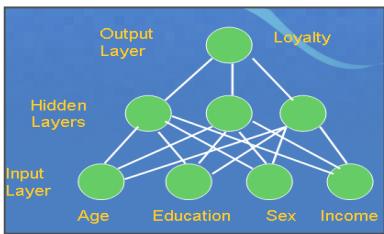
# SQL Server Data Mining Algorithms



- Association Rules
  - Supports market basket analysis to learn what products are purchased together



- Time Series
  - Forecasting algorithm used to predict future values from a time series



- Neural Net
  - Used for classification and regression tasks
  - Can explore extremely complex scenarios
  - Often challenging to configure and interpret its results



© 2014 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries. The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.