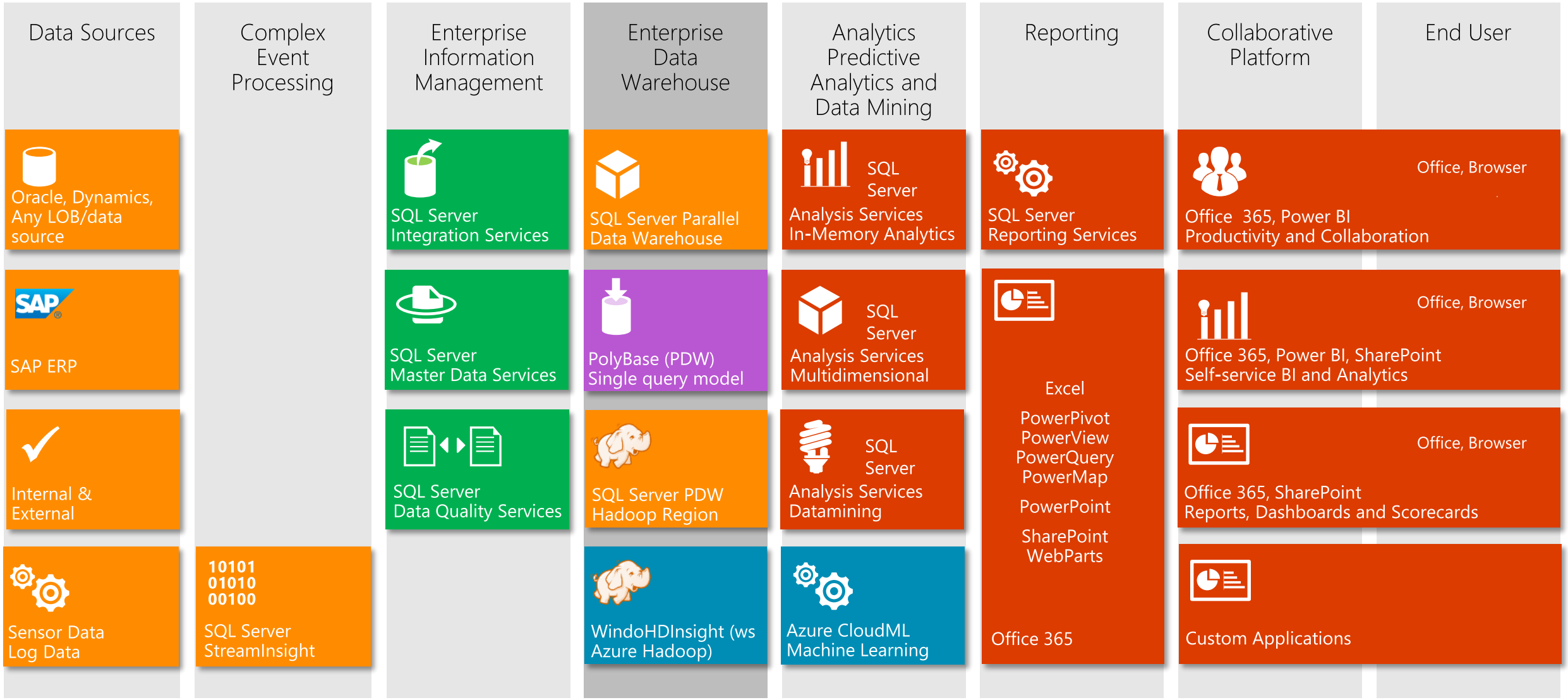
Microsoft Azure

Korkeakoulujen valtakunnalliset IT-päivät 2014
4.-5.11.2014 Marina Congress Center, Helsinki

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

Marko Hotti
Senior Solution Architect
Microsoft Data Platform
marko.hotti@microsoft.com
http://blogs.technet.com/markohot
@MarkoHotti

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



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

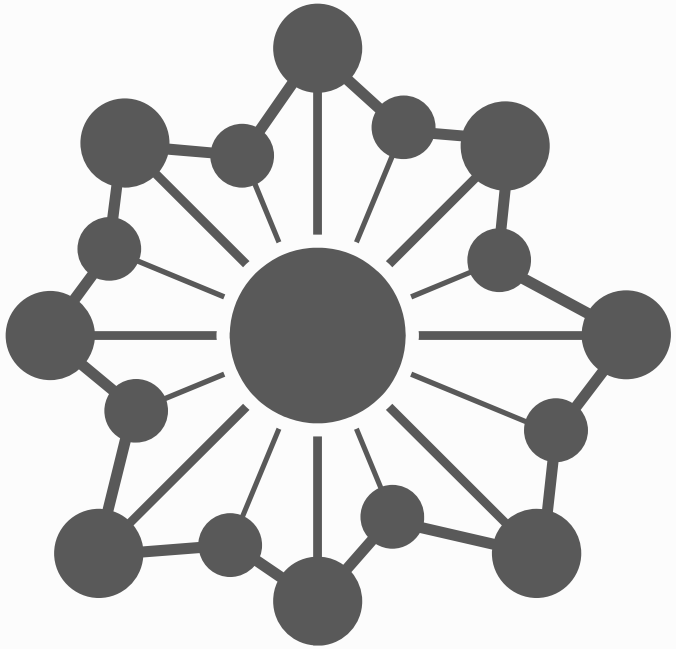
Financial and insurance risk calculation

Engineering modeling and simulation

Computational life sciences

Earth sciences

Data analytics



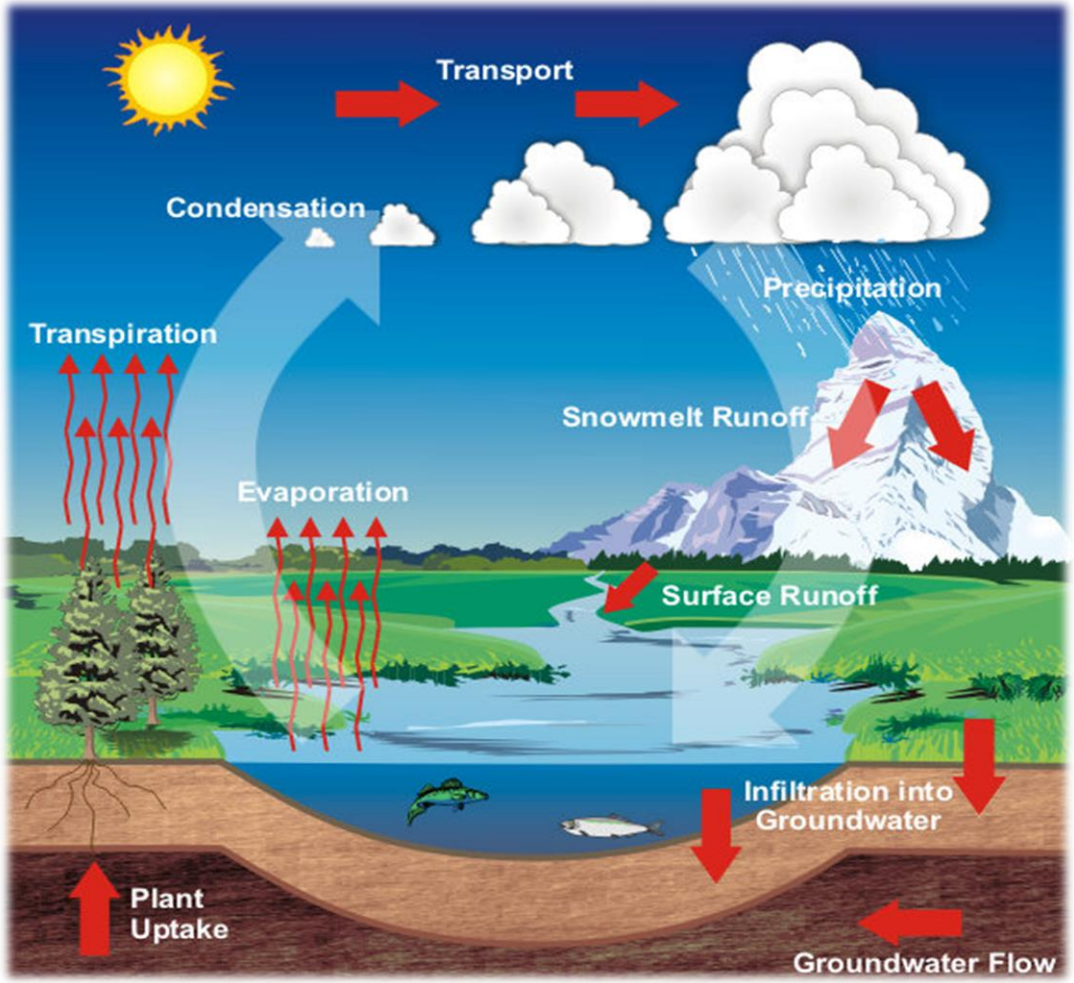
Azure Scenarios in Scientific Research

- 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

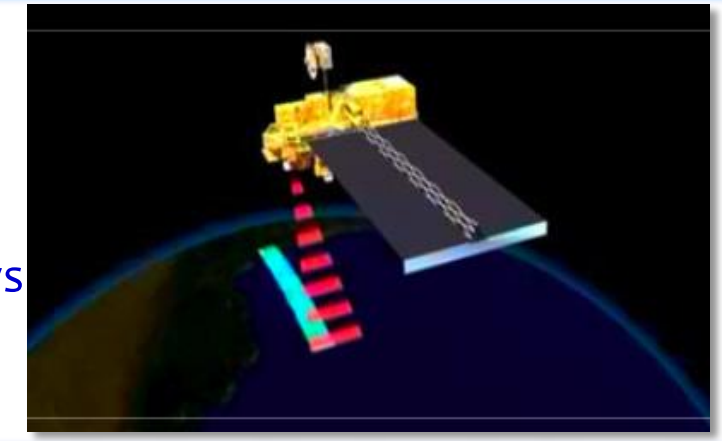
AzureMODIS – *Computing Evapotranspiration (ETP)*

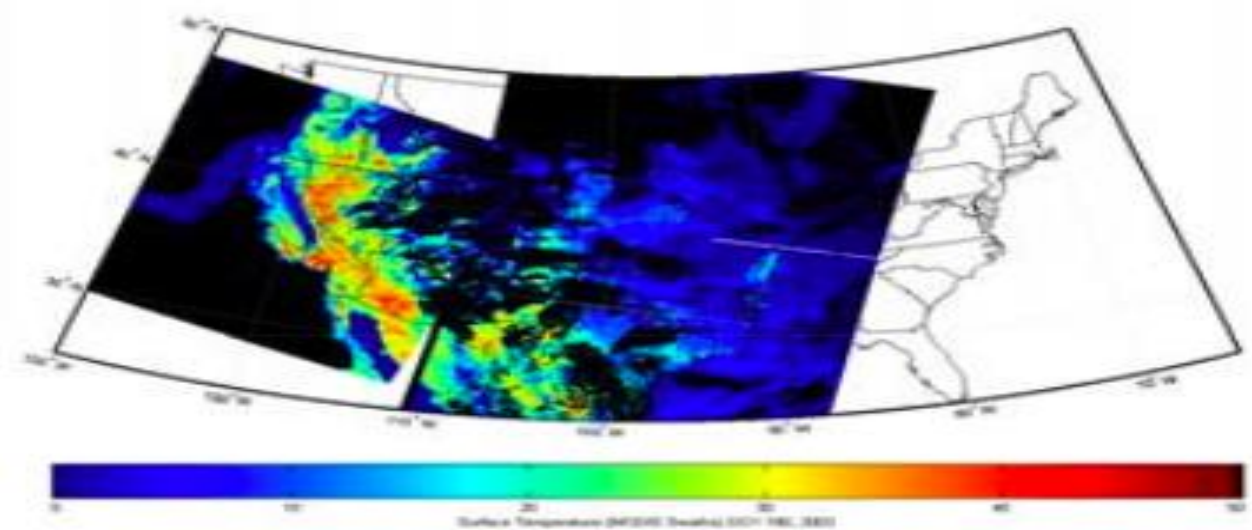
Catharine van Ingen (MSR), Jie Li, Marty Humphrey (UVA), Youngryel Ryu (UCB), Deb Agarwal (BWC/LBL), Keith Jackson (BL), Jay Borenstein (Stanford) , Team SICT: Vlad Andrei, Klaus Ganser, Samir Selman, Nandita Prabhu (Stanford), Team Nimbus: David Li, Sudarshan Rangarajan, Shantanu Kurhekar, Riddhi Mittal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



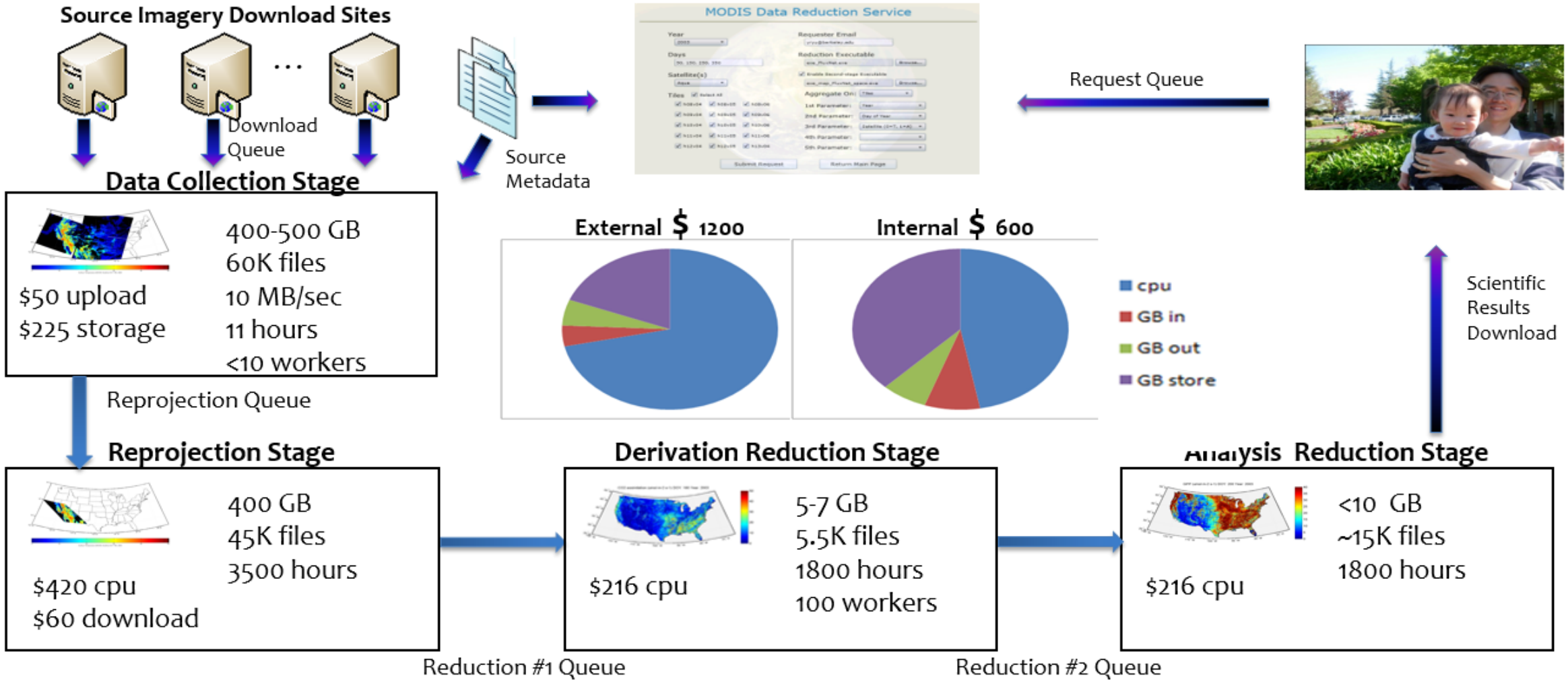
Two MODIS satellites

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



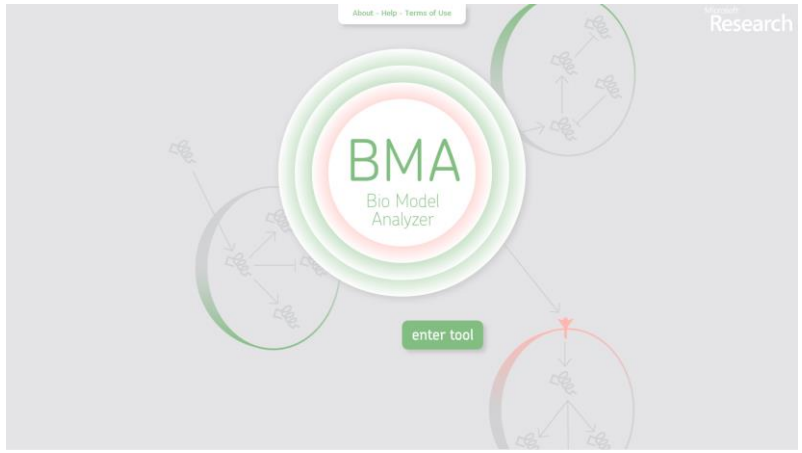


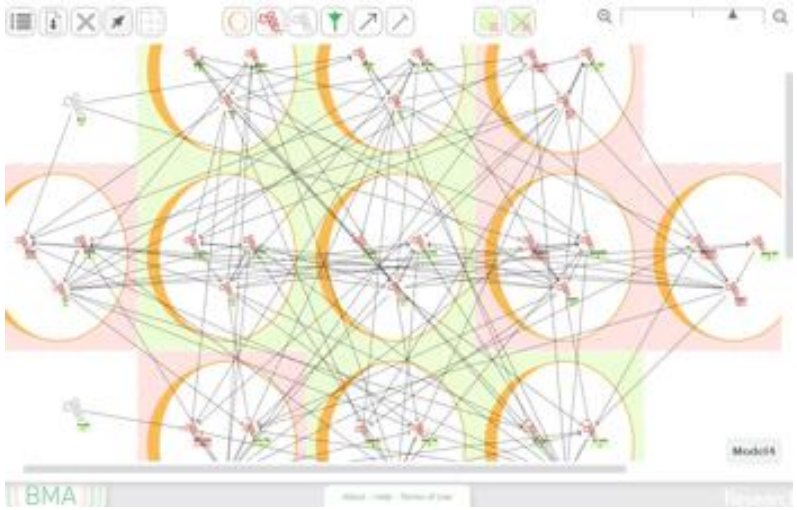
Computing Evapotranspiration for One US Year

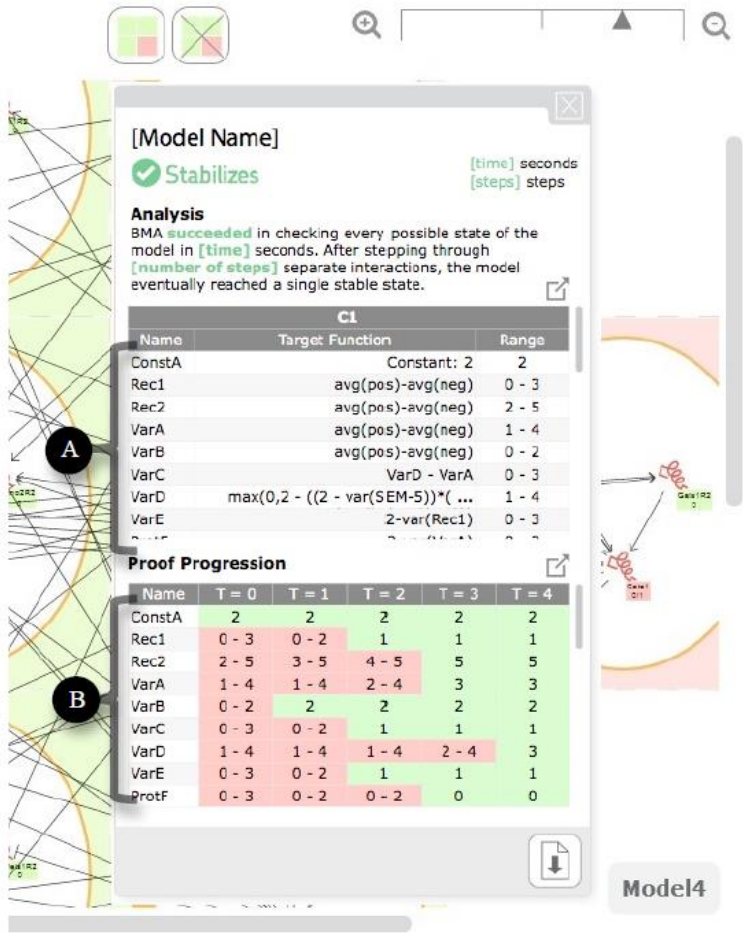


Total cost $1,800, where all storage costs assume 3 month project duration;

Computing Cancer



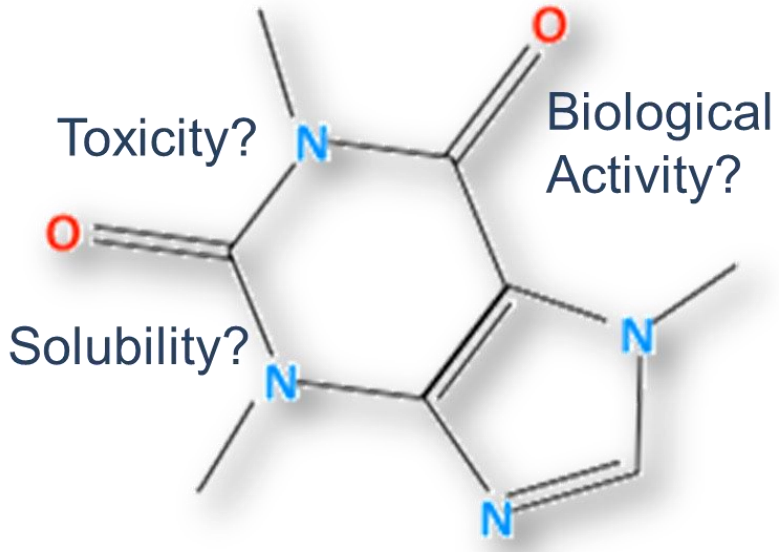




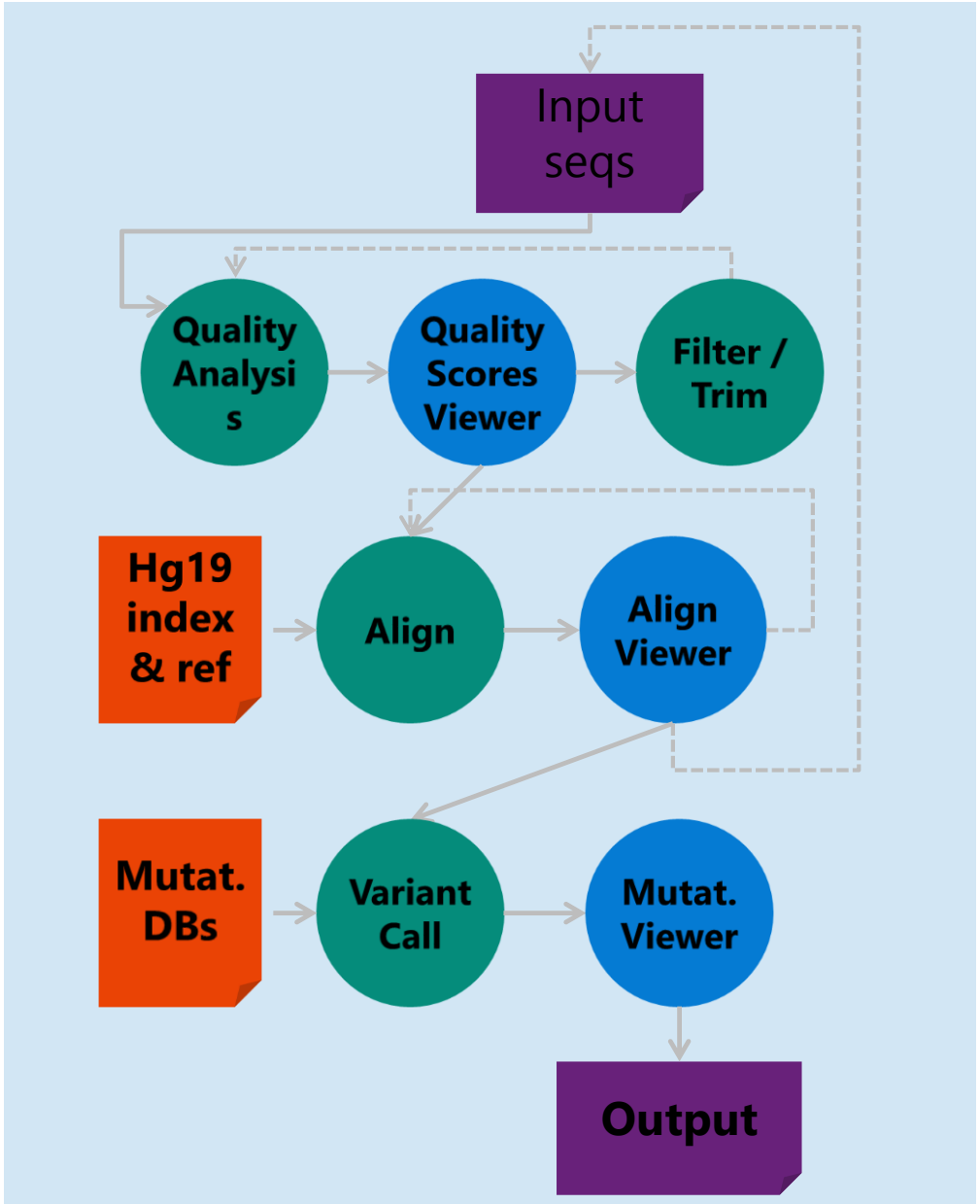
http://biomodelanalyzer.research.microsoft.com/

Jasmin Fisher, Microsoft Research Cambridge

Accelerating Drug Design - QSAR



NGS Pipelines in the Cloud



seqtk

seqtk

fastqc

Bowtie2

GATK

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

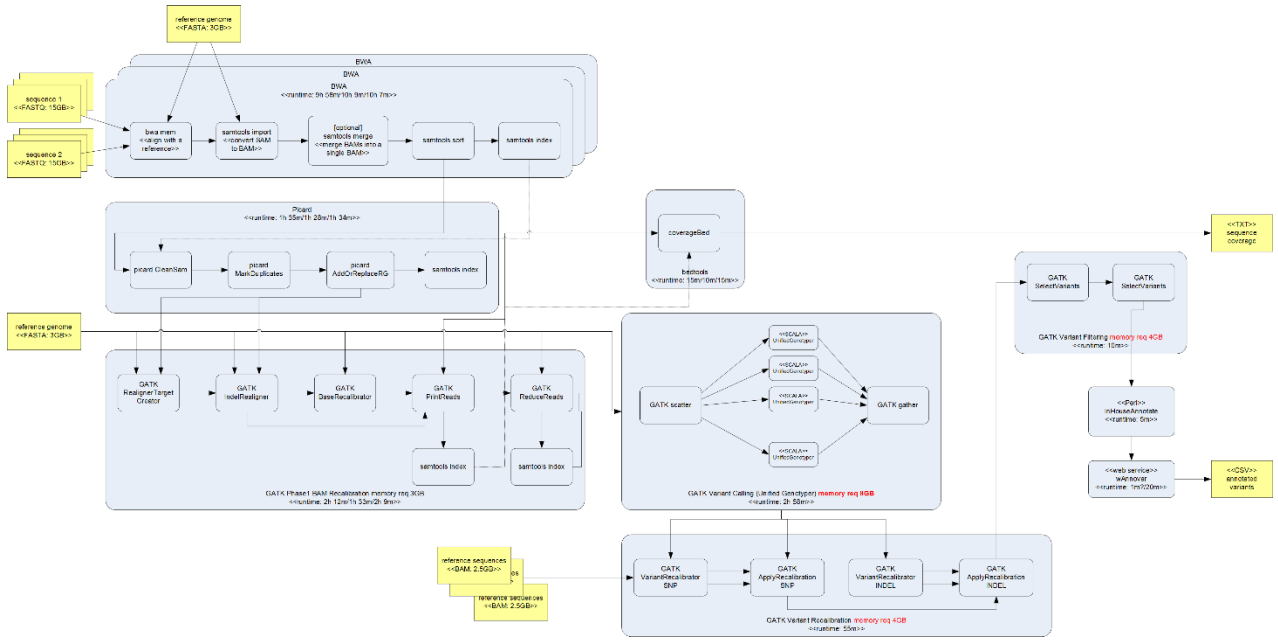
GenomeMaps

Cloud e-Genome

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

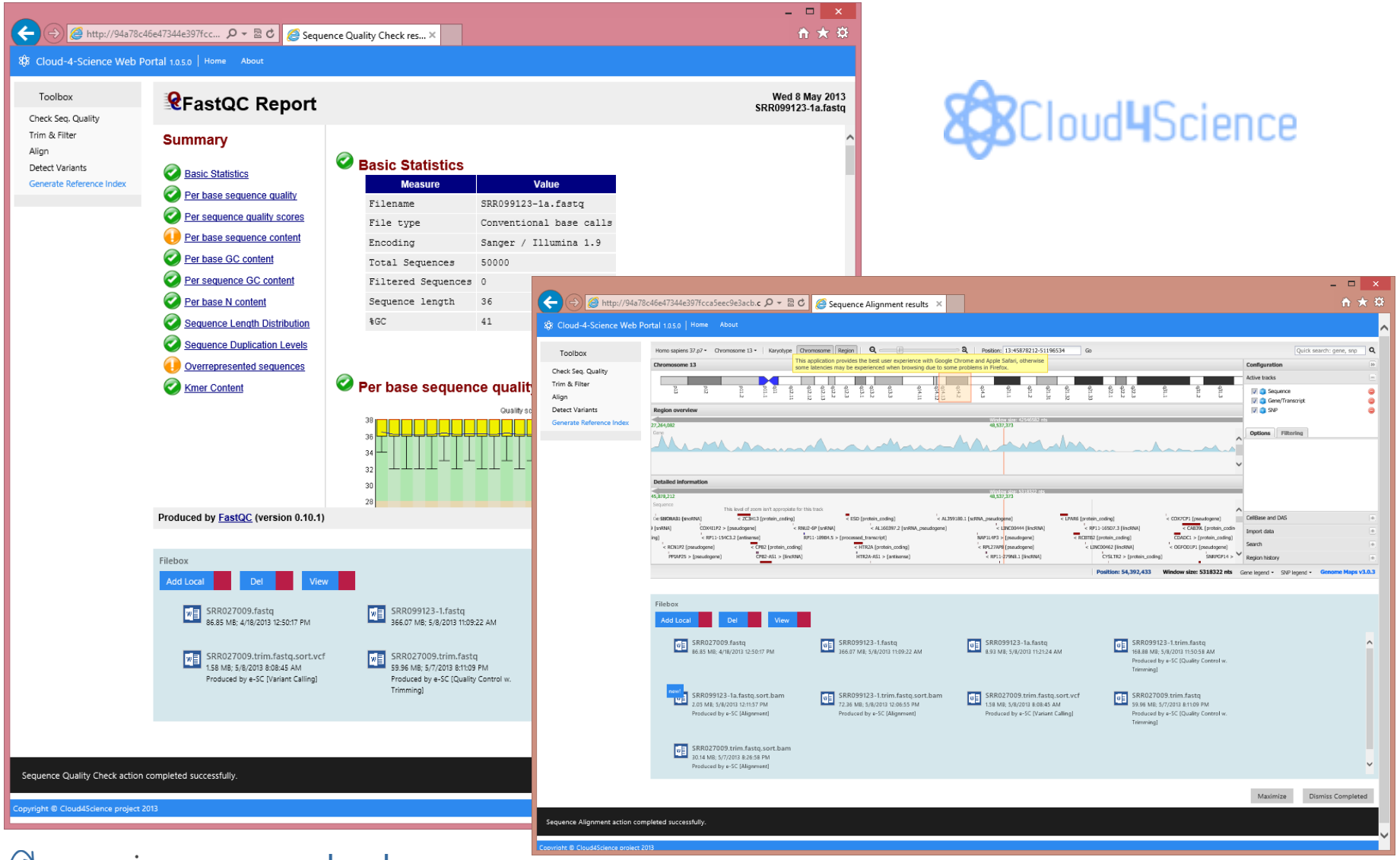
Objectives:

- 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



http://www.slideshare.net/pmissier/clinical-genomics-informatics-talk-lisbonpm

Cloud-enabled Genomics

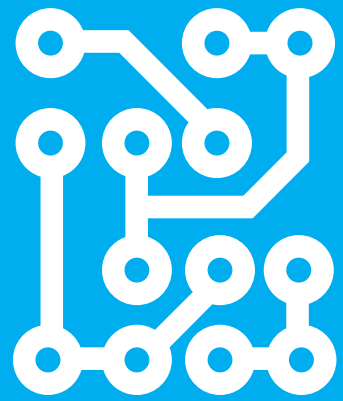


Cloud4Science

e-science central

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



Virtual machine operating system selection

- ALL
- PLATFORM IMAGES
- MY IMAGES
- MY DISKS

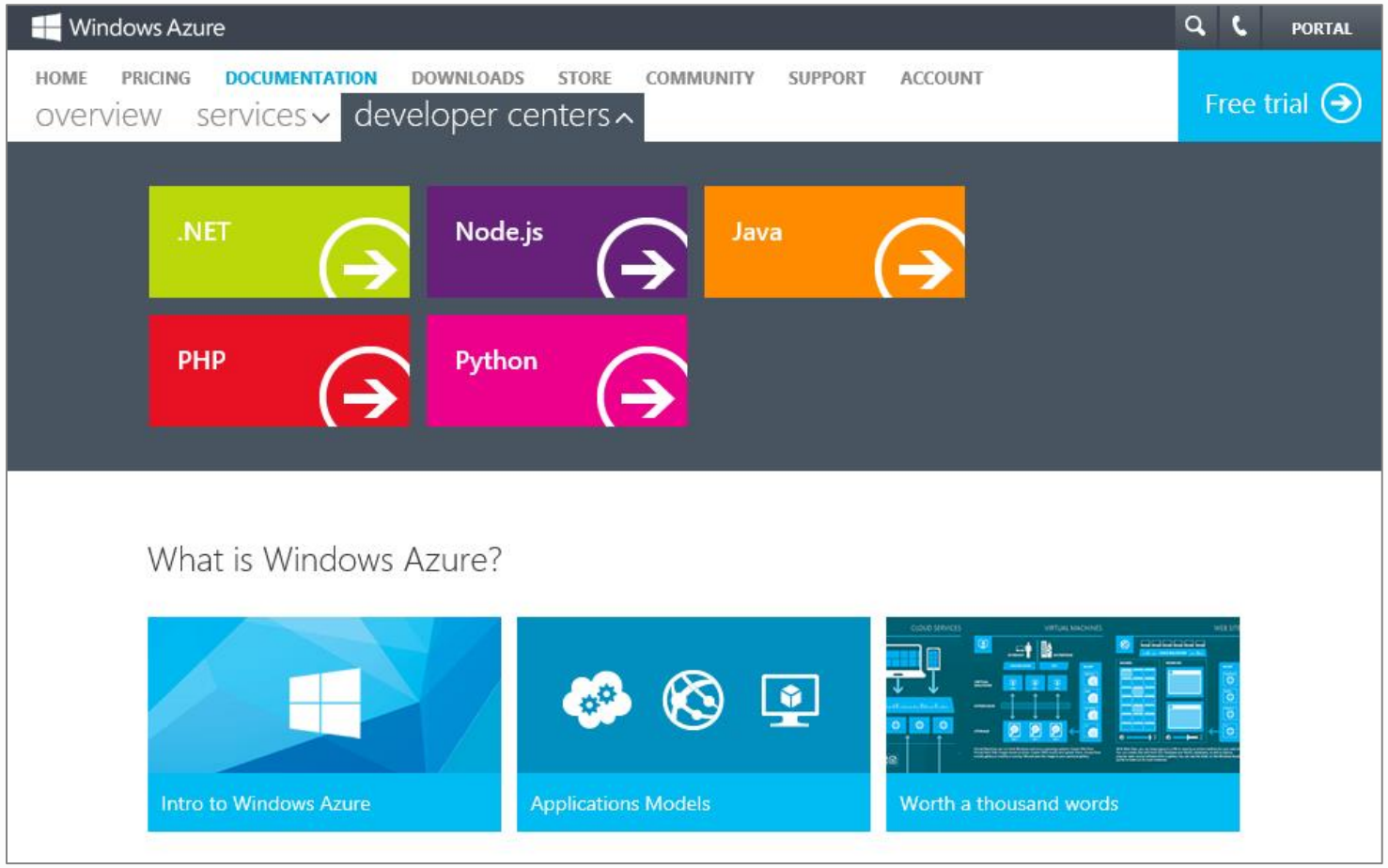
- Windows Server 2008 R2 SP1
- Windows Server 2012 Datacenter
- OpenLogic CentOS 6.3
- openSUSE 12.3
- RightScale Linux v13
- SUSE Linux Enterprise Server 11 SP2
- Ubuntu 12.04
- Ubuntu 12.10
- wenmingsaved
- whitehall
- boothdemo1-boothdemo1-0-2012

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.

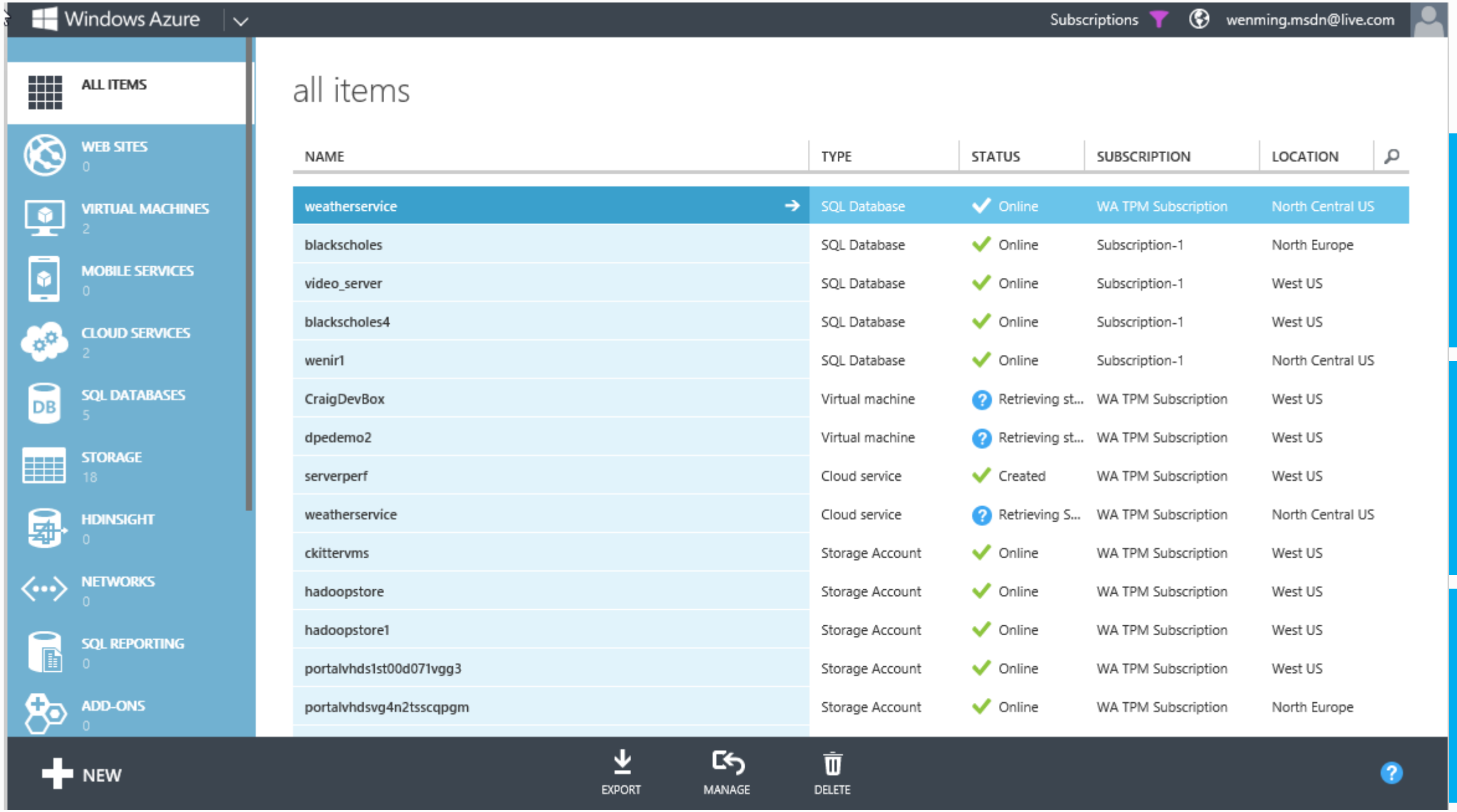
PUBLISHER	Microsoft SQL Server Group
OS FAMILY	Windows
LOCATION	East Asia;Southeast Asia;North Europe;West Europe;East US;West US

2 3 4



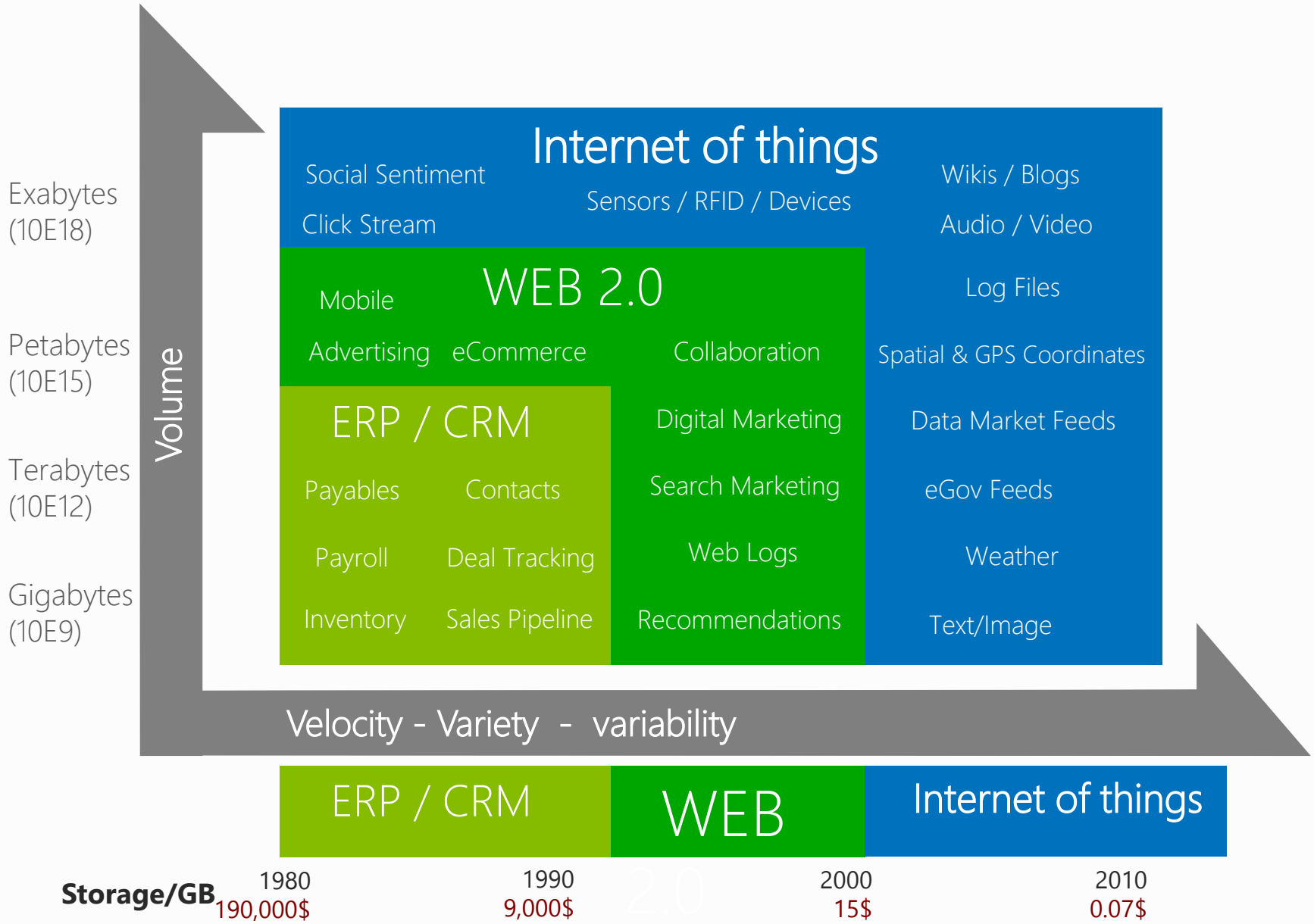
→ http://WindowsAzure.com

Multiple/Any languages (Fortran too)



Microsoft Azure

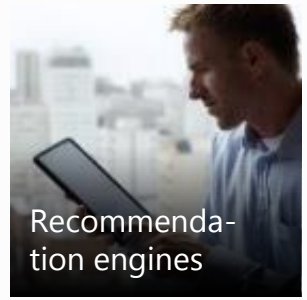
Big Data

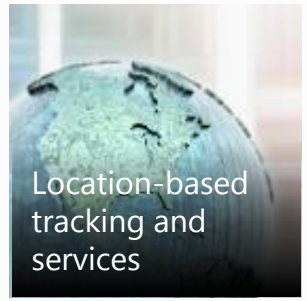

Exabytes (10E18)
Petabytes (10E15)
Terabytes (10E12)
Gigabytes (10E9)
Volume
Internet of things
Social Sentiment
Click Stream
Sensors / RFID / Devices
Wikis / Blogs
Audio / Video
Log Files
Spatial & GPS Coordinates
Data Market Feeds
eGov Feeds
Weather
Text/Image
WEB 2.0
Mobile
Advertising
eCommerce
Collaboration
Digital Marketing
Search Marketing
Web Logs
Recommendations
ERP / CRM
Payables
Contacts
Payroll
Deal Tracking
Inventory
Sales Pipeline
Velocity - Variety - variability
ERP / CRM
WEB
Internet of things
Storage/GB
1980
190,000$
1990
9,000$
2000
15$
2010
0.07$

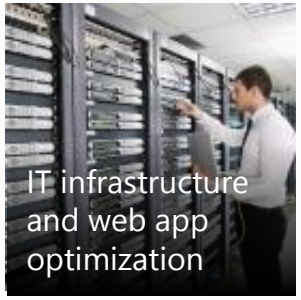
Predictive Analytics

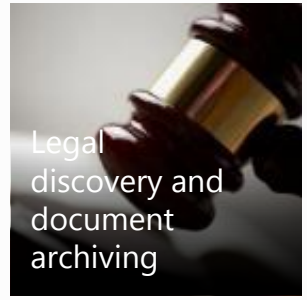
Predicting future performance from historical data

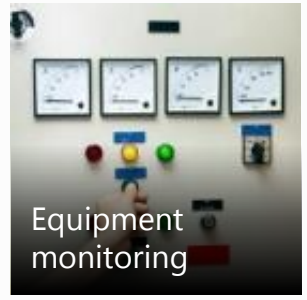
Predictive analytics should address the likelihood of something happening in the future, even if it is just an instant later*

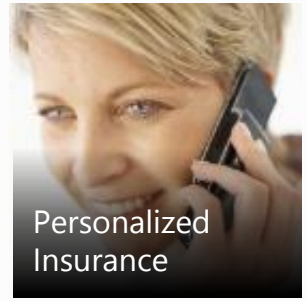












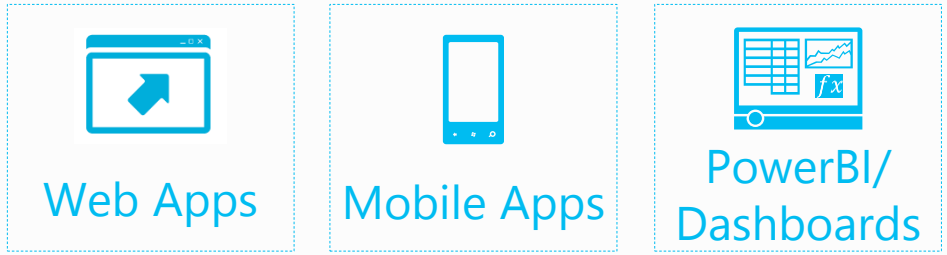
“Azure Machine Learning offers a data science experience that is directly accessible to business analysts and domain experts, reducing complexity and broadening participation through better tooling.”

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

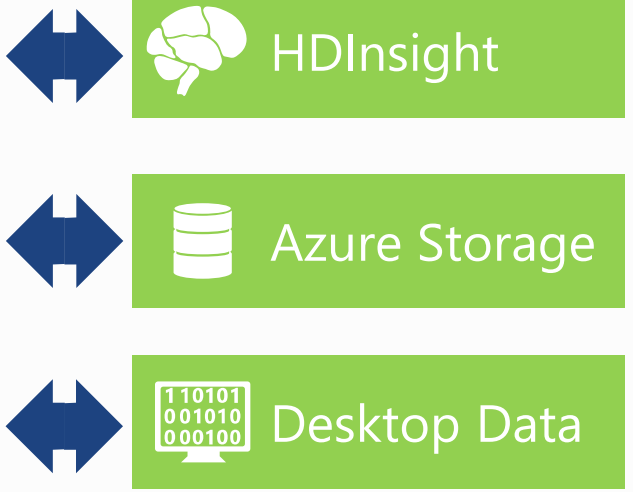
- Tested models available as an url that can be called from any end point

Azure Portal & ML API service
and the Azure Ops Team

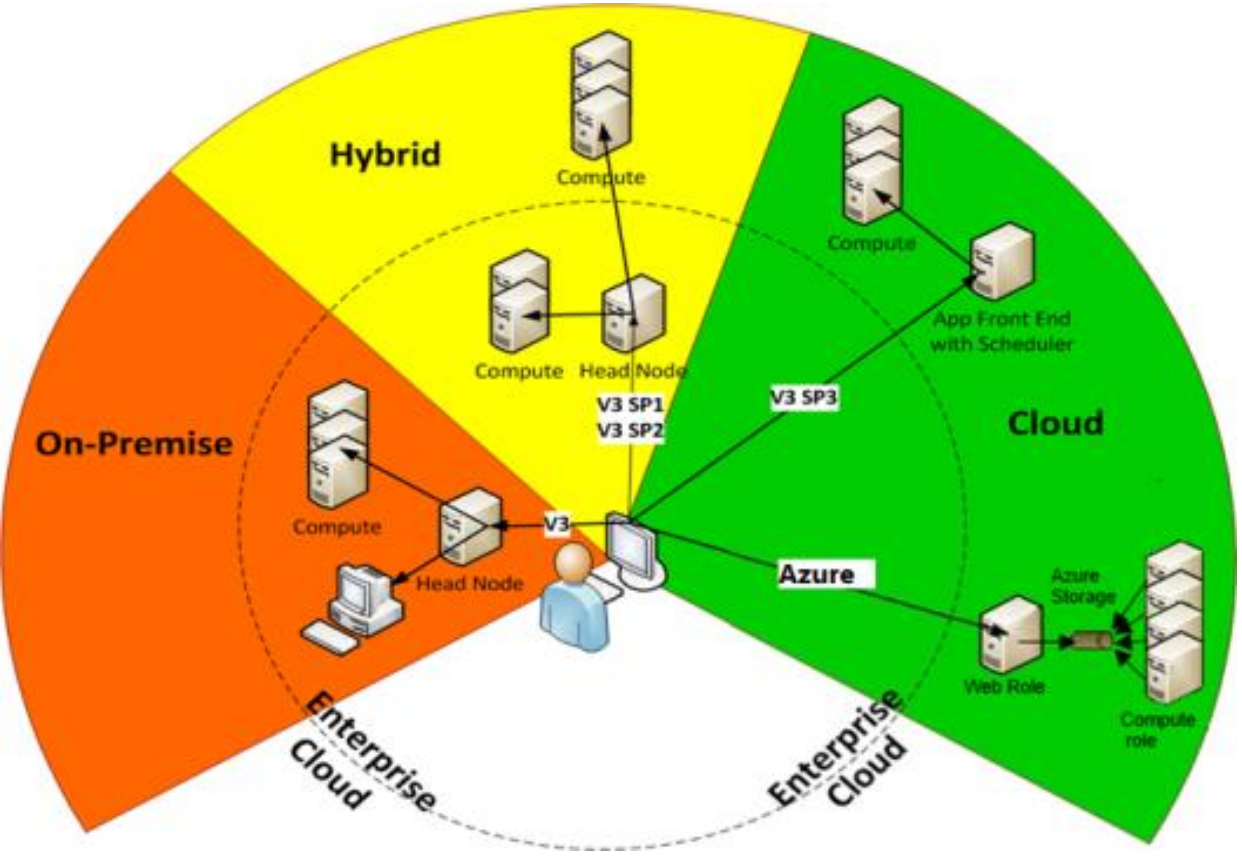
- Create ML Studio workspace
- Assign storage account(s)
- Monitor ML consumption
- See alerts when model is ready
- Deploy models to web service

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



Microsoft HPC Scenarios



On-Premises

Your own servers inside your enterprise cloud.

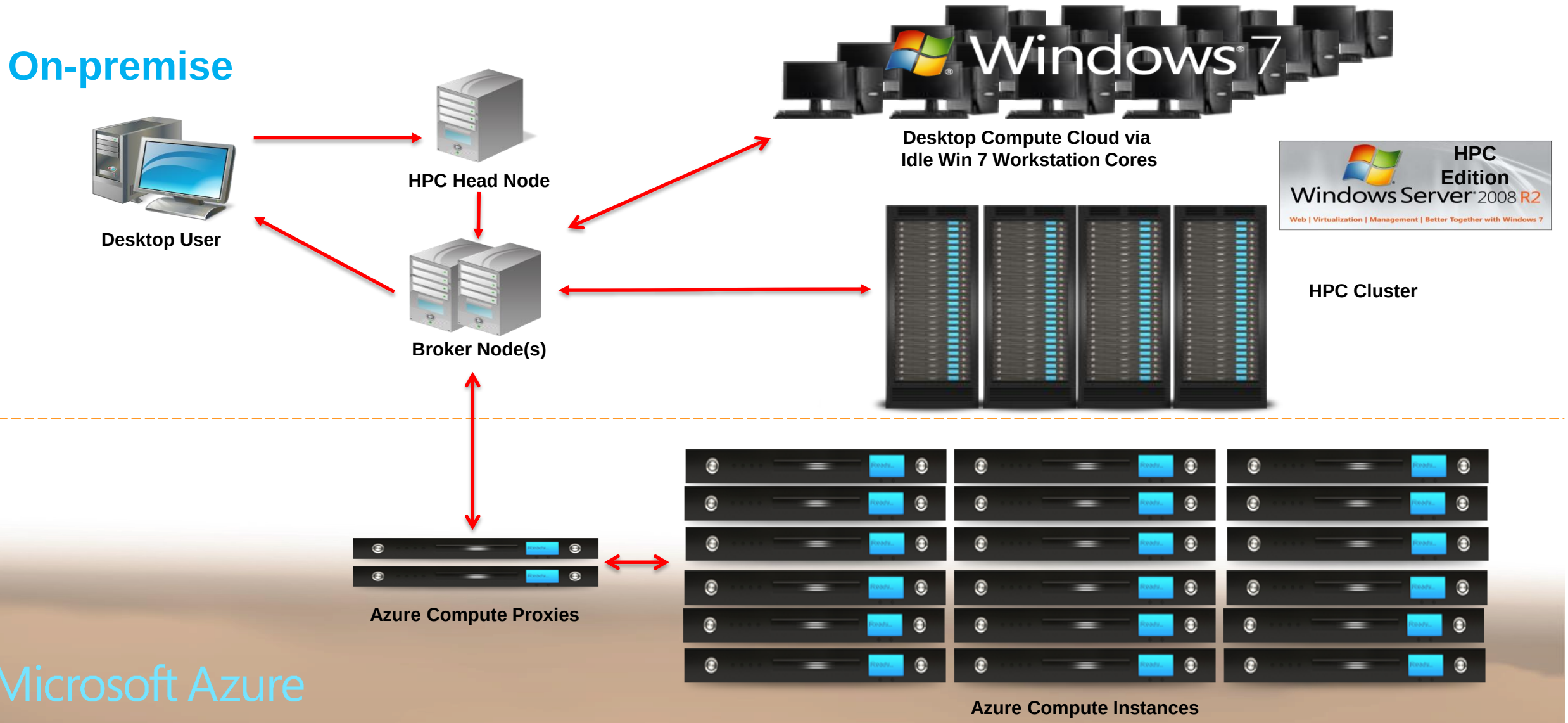
Hybrid

Burst from enterprise to cloud.

Cloud

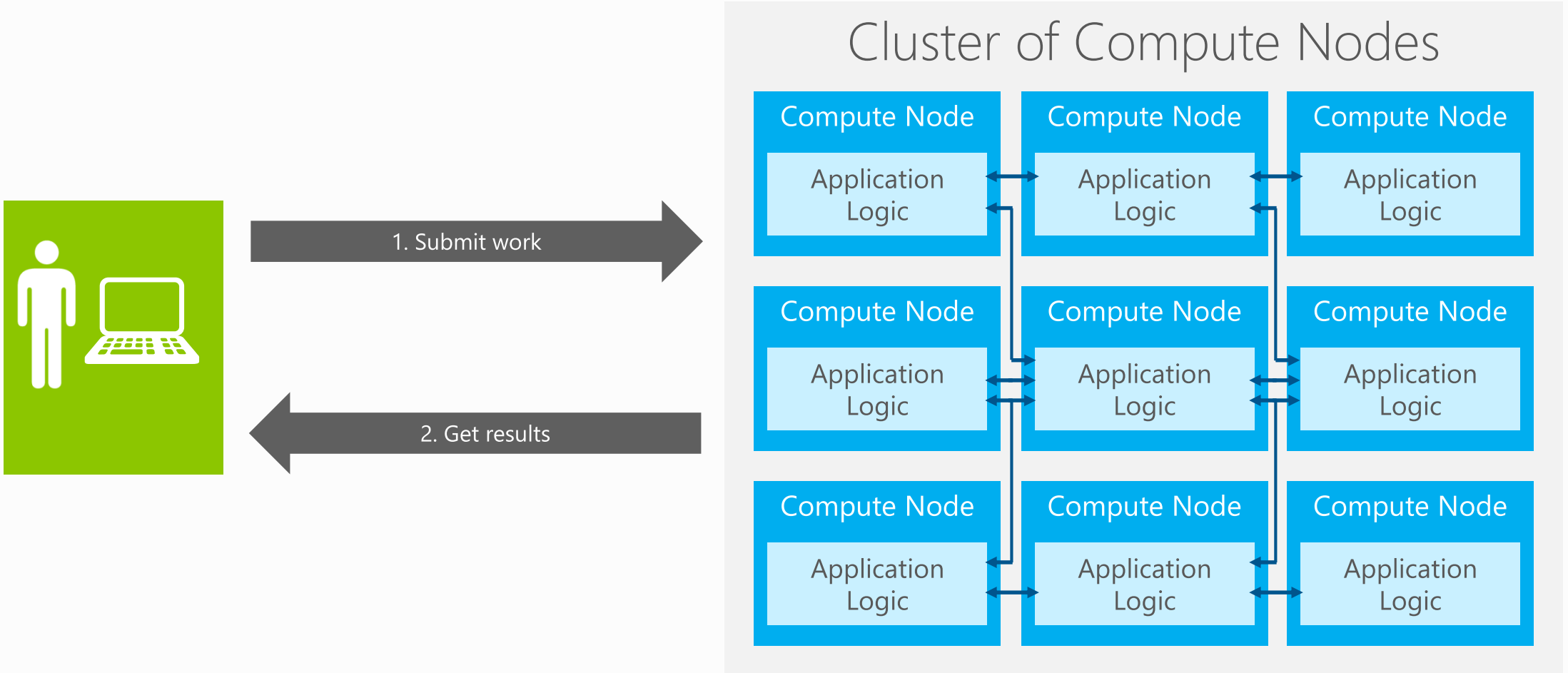
Head node in cloud or compute in cloud
Or both.

HPC Components (On Premises View)



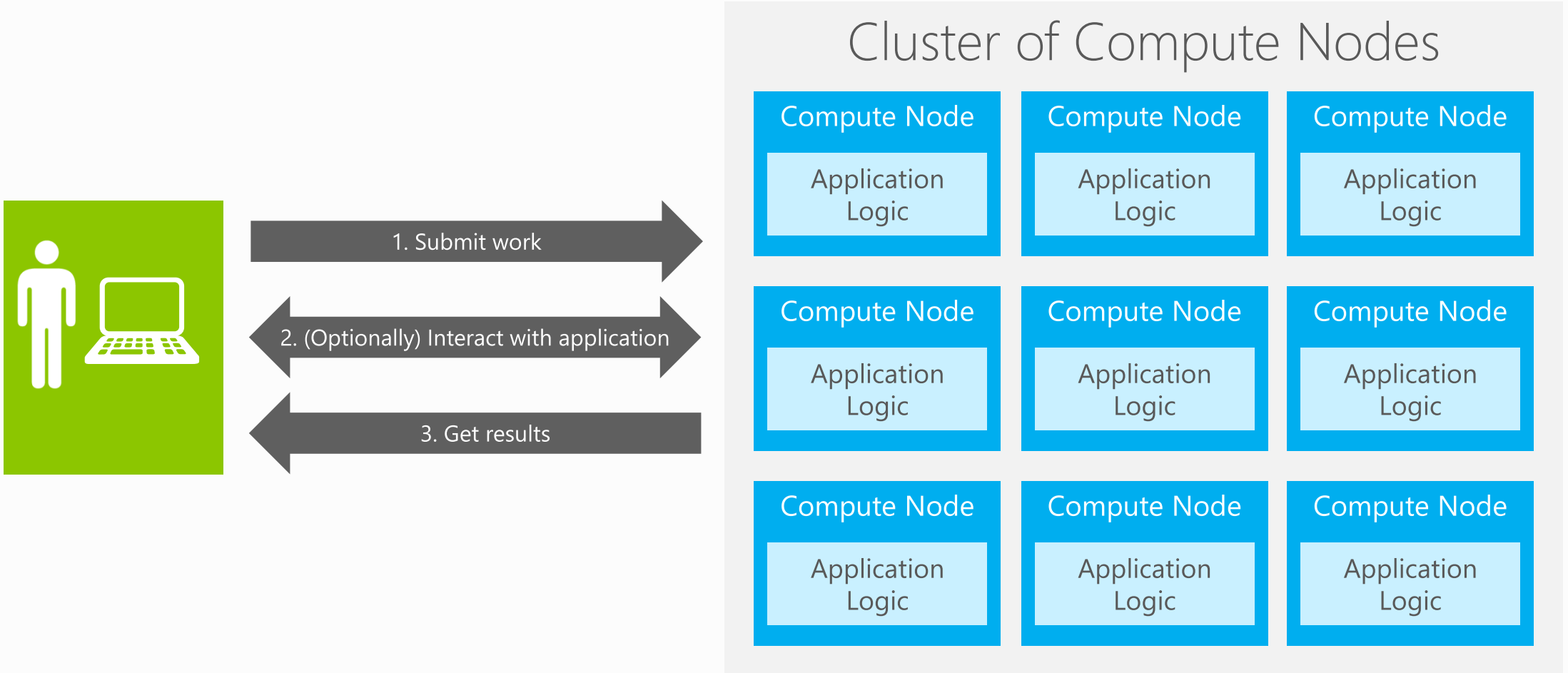
Tightly Coupled App (MPI)

An illustration



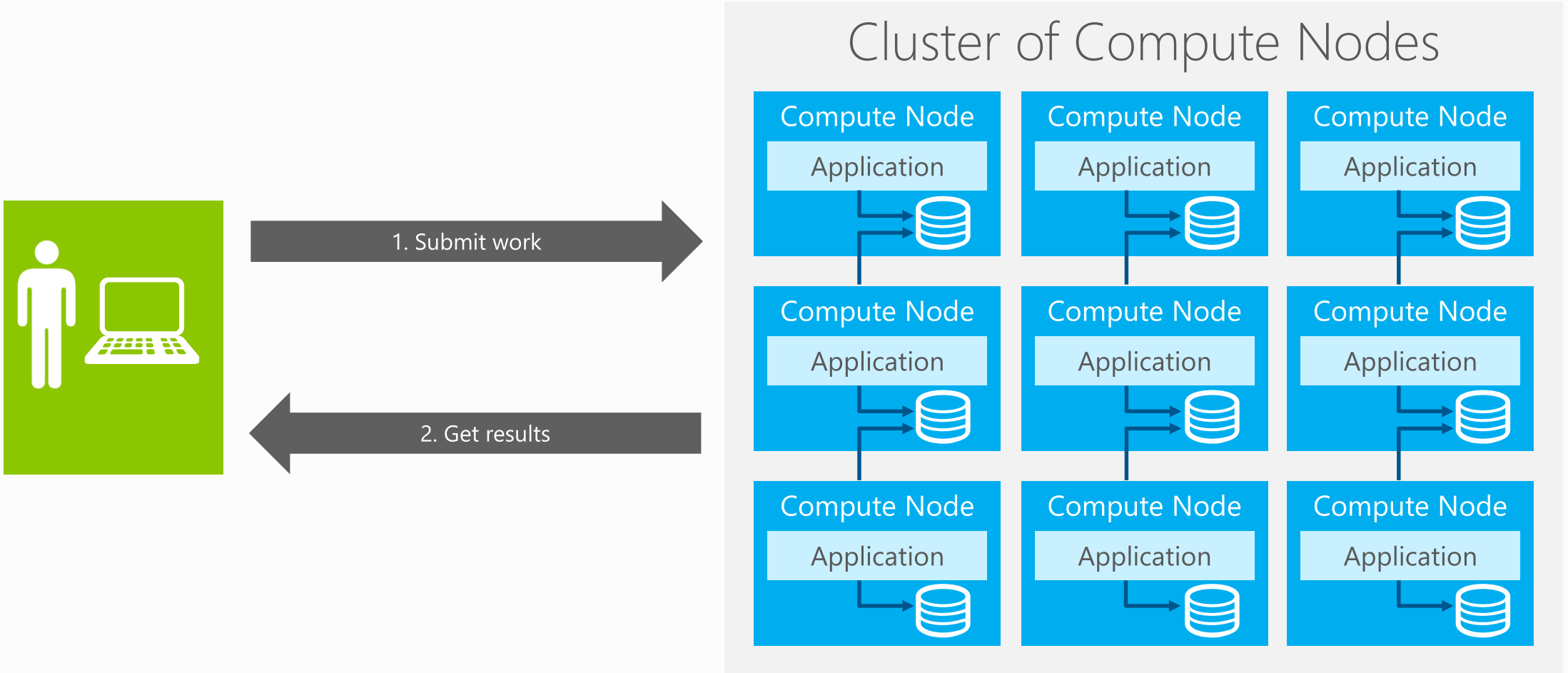
Embarrassingly Parallel Apps

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

Insights

Integrate with Microsoft business intelligence

Enterprise-ready

Choice of deployment on Windows Server + Windows Azure

Integrate with Windows components (Active Directory, Microsoft System Center)

Broader access

Easy installation and configuration of Hadoop on Windows

Simplified programming with Microsoft .NET and JavaScript integration

Integrate with SQL Server Data Warehousing

Contributions proposed back to community distribution

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

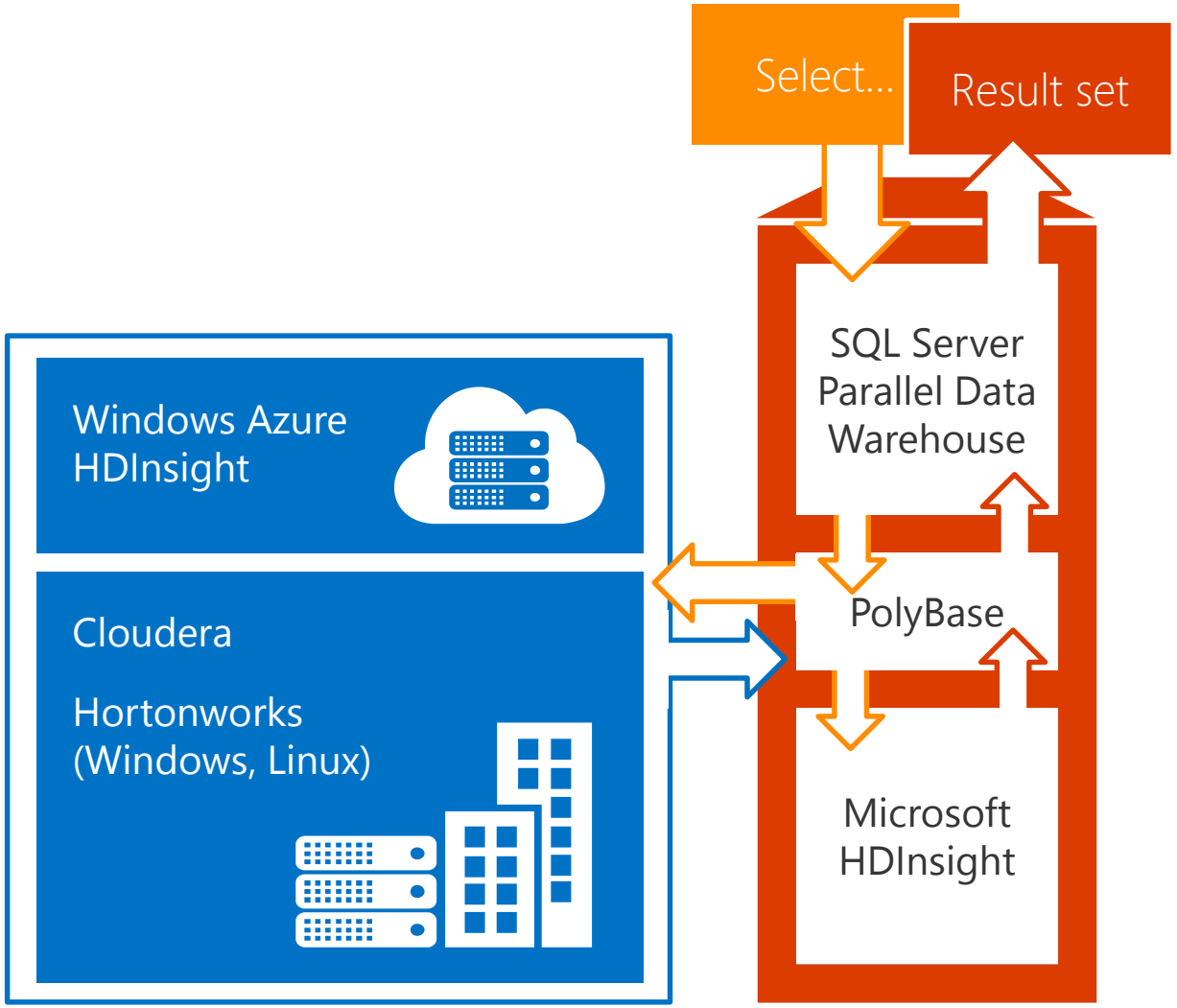
- Near real-time performance with In-Memory
- Scale-out to accommodate your growing data
- Remove DW bottlenecks with MPP SQL Server
- Concurrency that fuels rapid adoption

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 or 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)

Microsoft

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

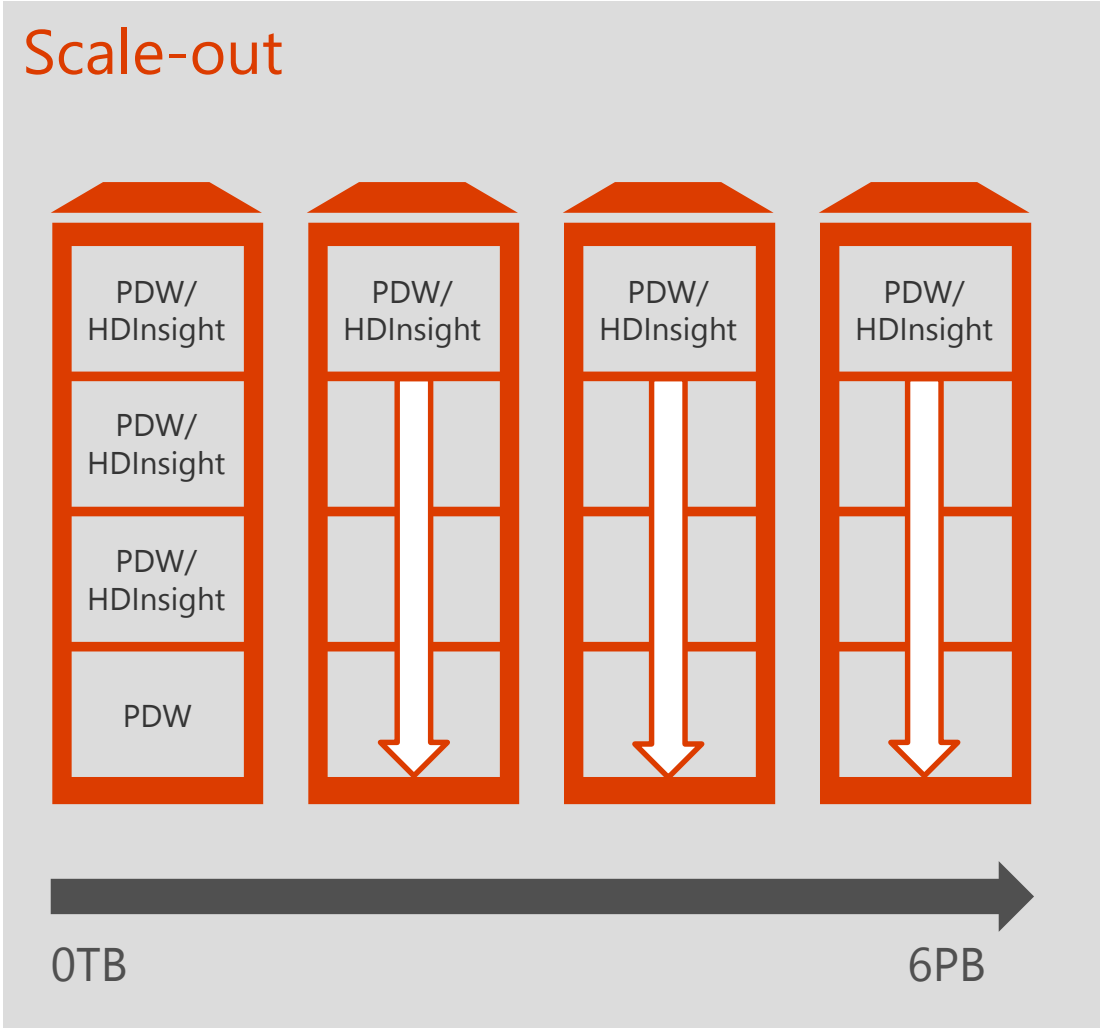
- Parallelized CREATE EXTERNAL TABLE AS SELECT (CETAS)
- External table as the destination; creates a set of HDFS files

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

Microsoft

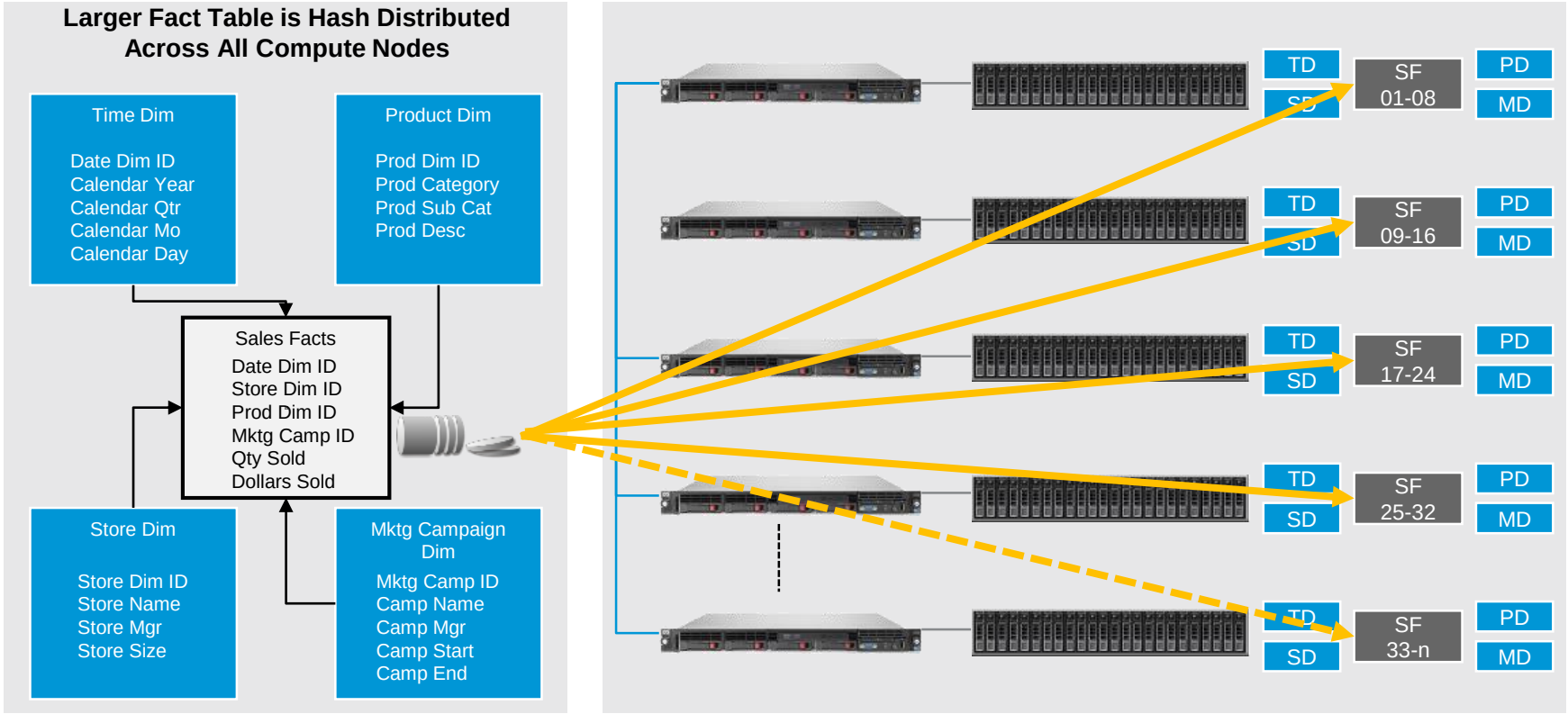
Scaling out relational data to petabytes

Scale-out technologies in the Analytics Platform System



- 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

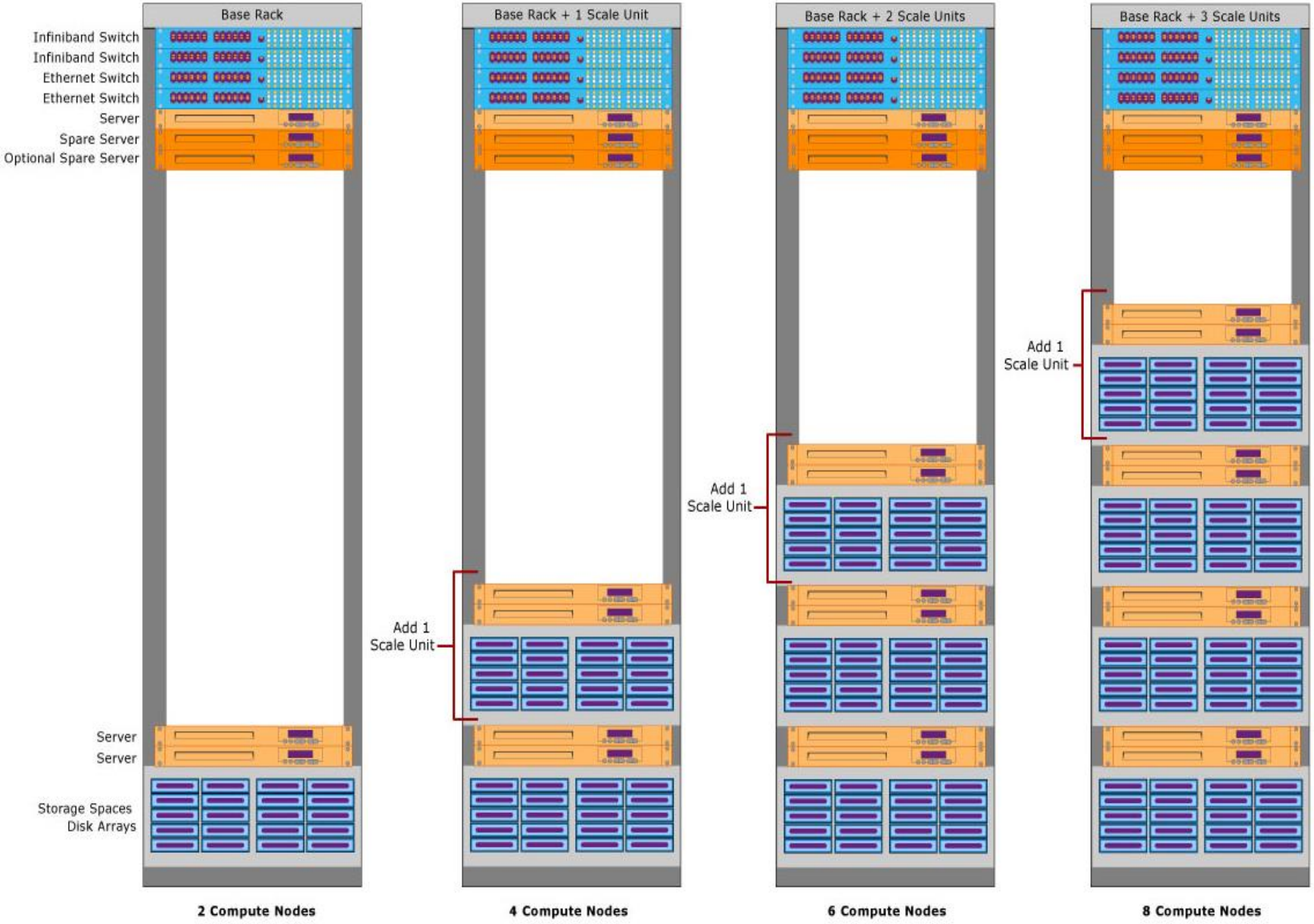
Ultra Shared Nothing architecture: Distribution



Microsoft

HP Base Rack

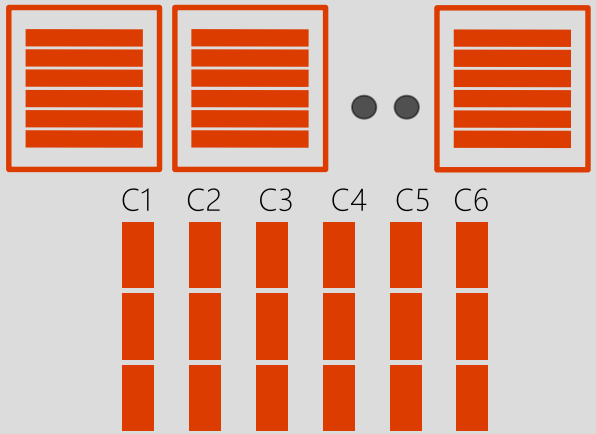
The following diagram shows the architecture for an HP Base rack. This Base rack ships with 2 Compute nodes which have the same architecture as a Scale unit. To add more Compute nodes to the Base rack, you can add Scale units. A full HP Base rack has 8 Compute nodes.



Blazing fast performance

MPP and In-memory columnstore for next-generation performance

Columnstore index representation

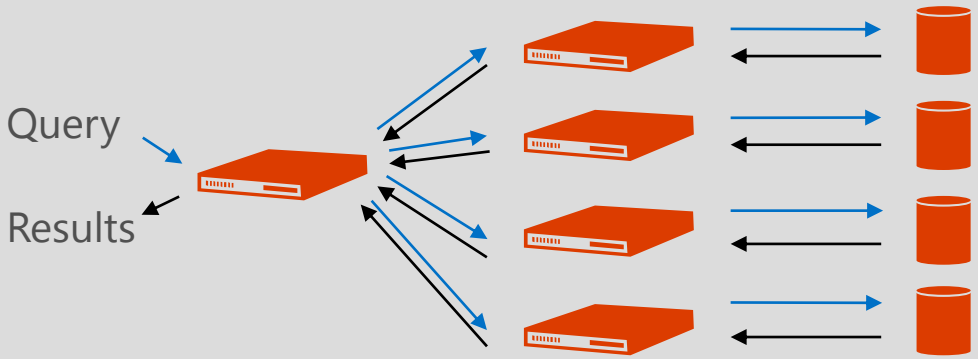


Up to 100x faster queries

Up to 15x more compression

Updatable clustered columnstore vs. table with customary indexing

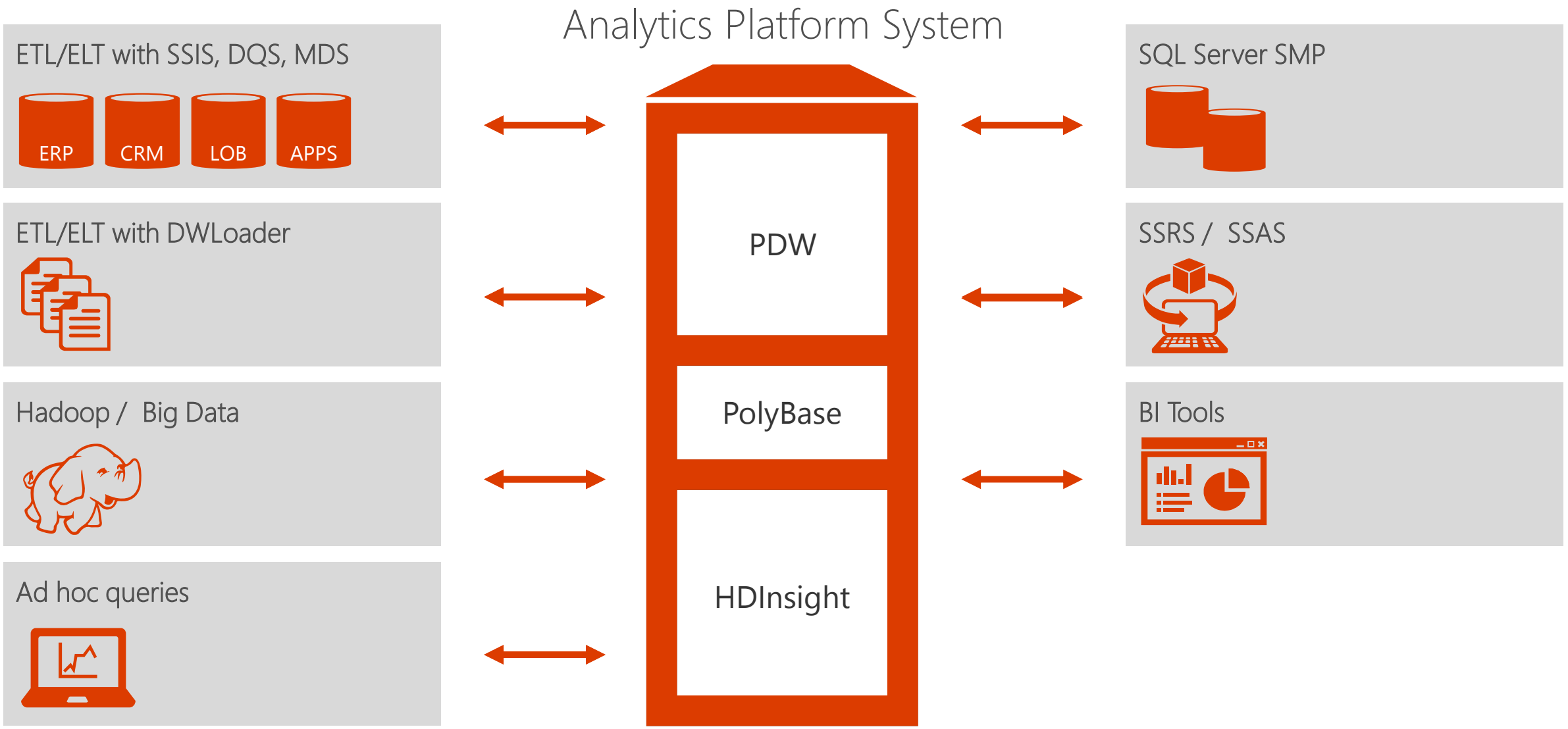
Parallel query execution



- 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

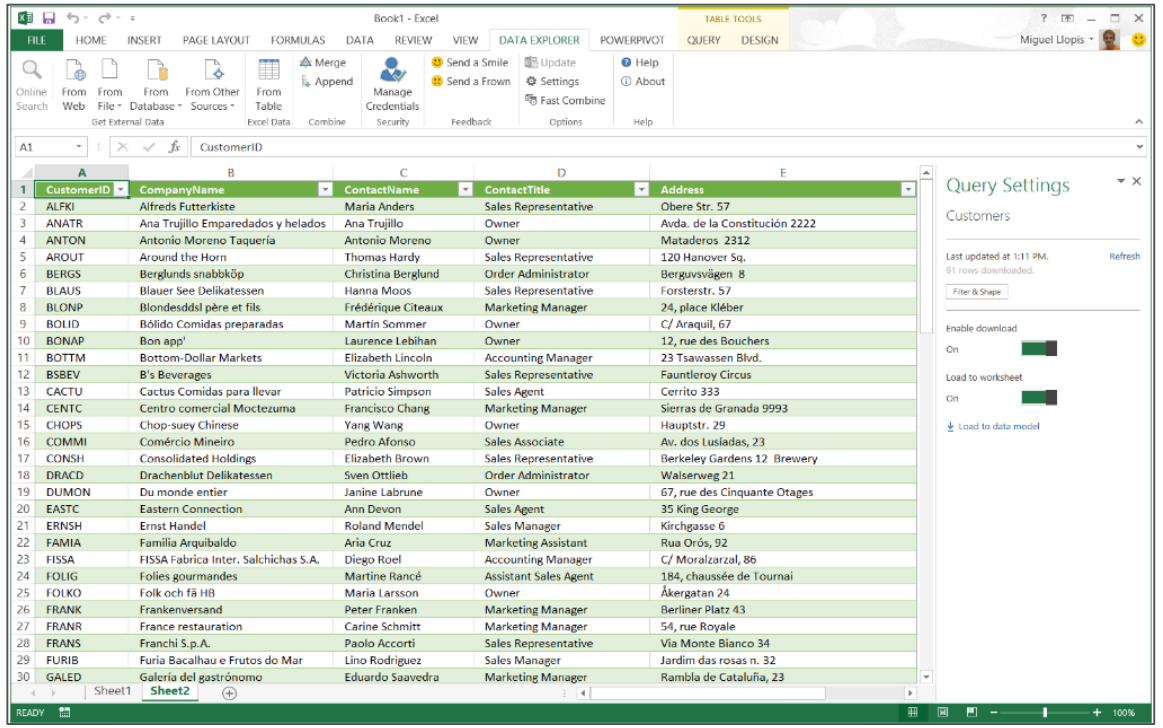
Great performance with mixed workloads



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



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

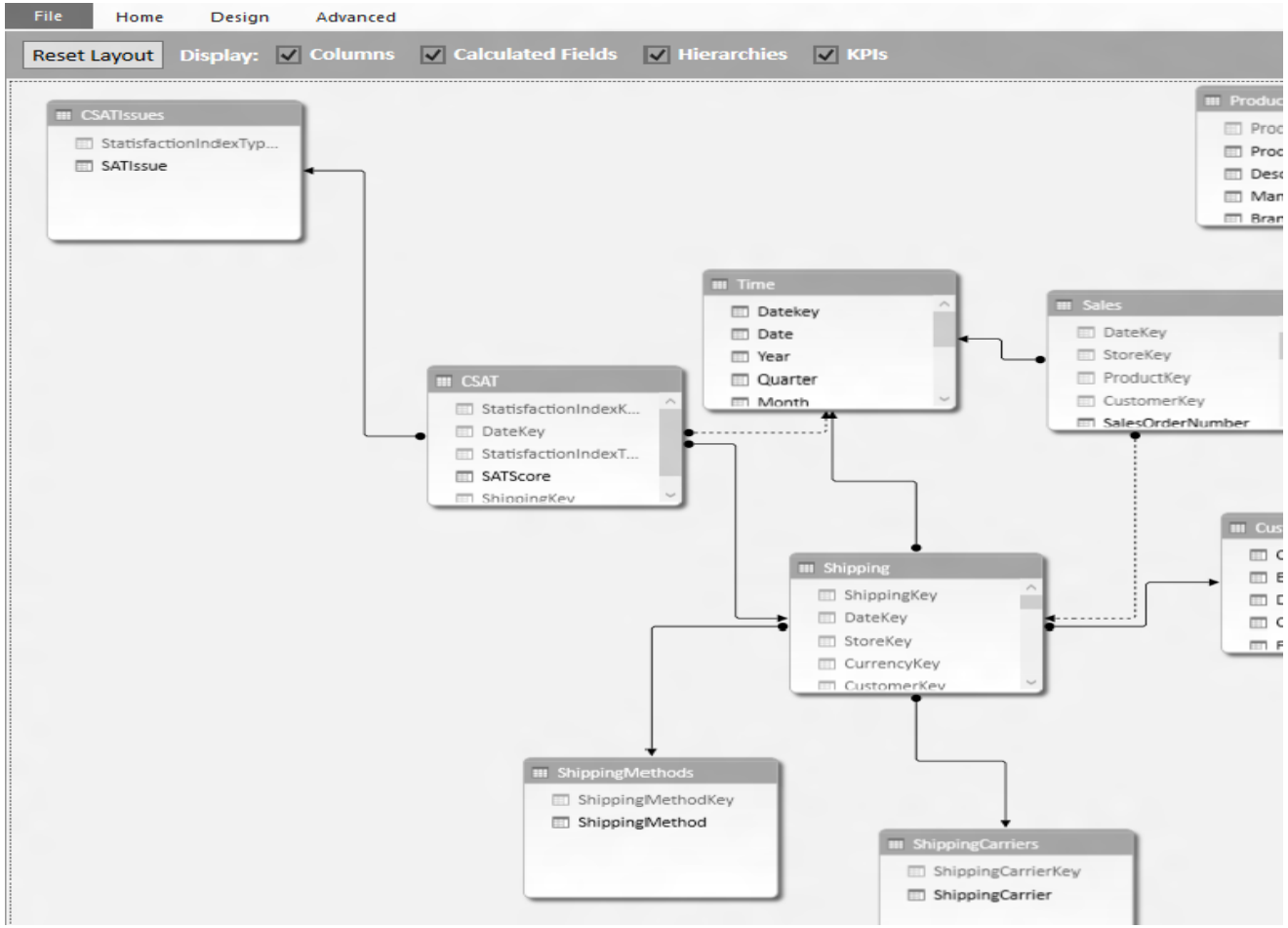
Model relationships, custom measures, hierarchies, and KPI's

Visualize & Explore

Bring your data to life with interactive visualization

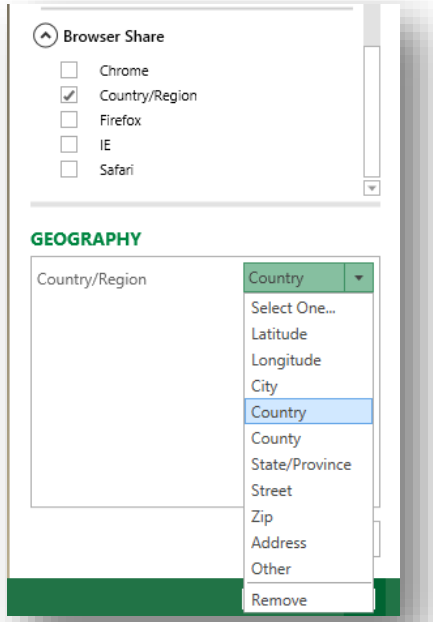
Explore data in new ways to discover hidden insights

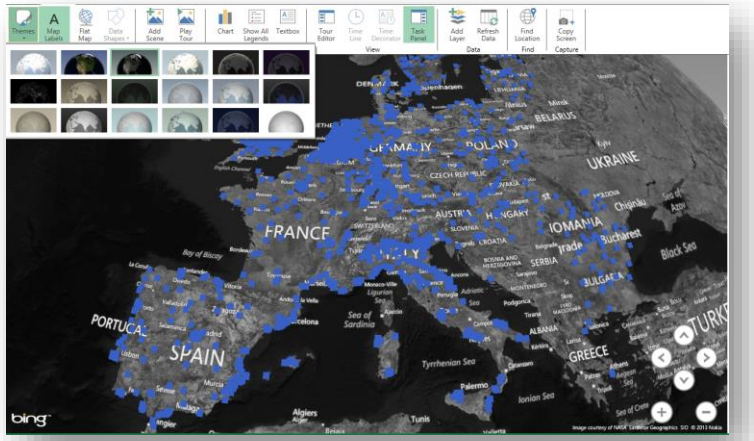
PowerPivot

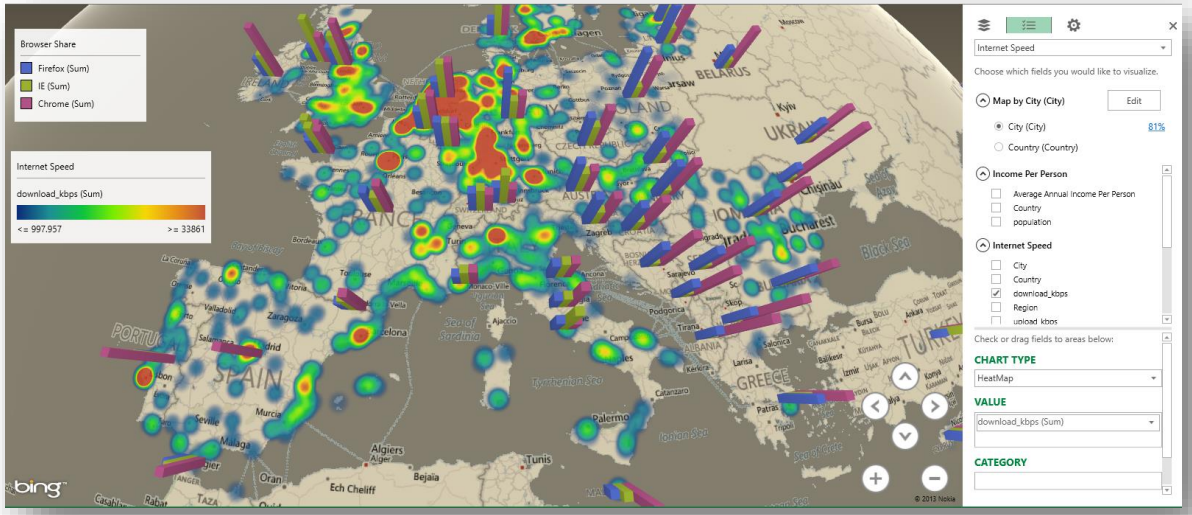


Map Data

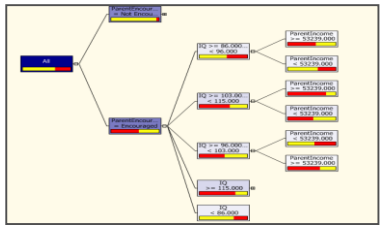
- Geospatial and temporal data in Excel
  - Calculated Fields and Hidden Columns
- Geo-code and Themed maps with Bing
- Visualize multiple layers at once
- 3D Columns, Bubble/Pies, Heat Maps
  - Regions for geopolitical boundaries



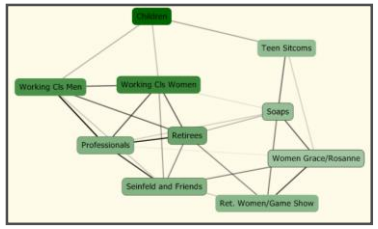




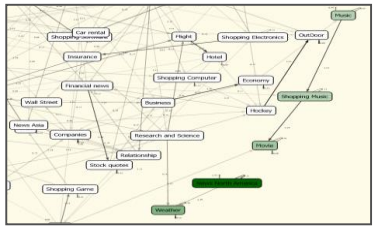
SQL Server Data Mining Algorithms



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

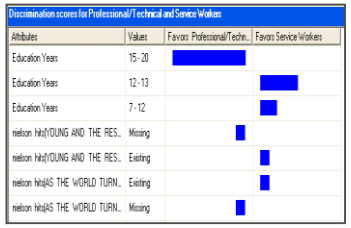


- Clustering
  - Finds natural groupings inside data

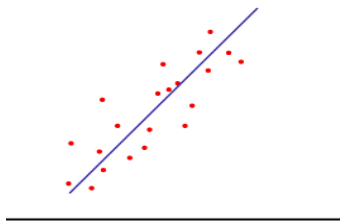


- Sequence Clustering
  - Groups a sequence of discrete events into natural groups based on similarity
  - Use this algorithm to understand how visitors use your Web site

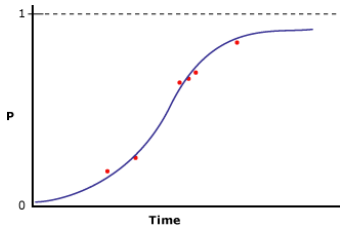
SQL Server Data Mining Algorithms



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

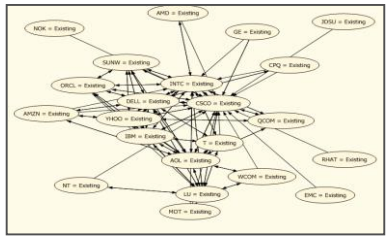


- Linear Regression
  - Finds the best possible straight line through a series of points
  - Used for prediction analysis

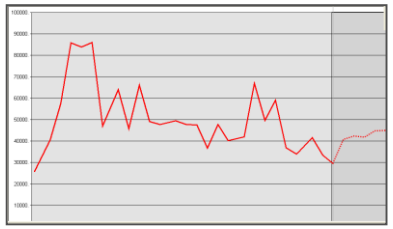


- Logistic Regression
  - Fits to an exponential factor
  - Used for prediction analysis

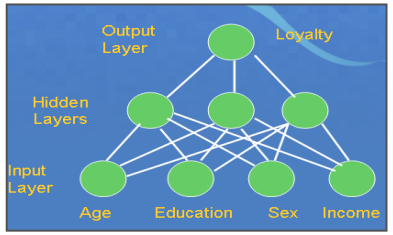
SQL Server Data Mining Algorithms



- 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