



SQL Server 2014

Faster Insights from Any Data

(Level 300 Deck)



SQL Server 2014 Faster Insights from Any Data

(Level 300 Deck)





SQL Server 2014 Faster Insights from Any Data

(Level 300 Deck)



SQL Server 2014 Faster Insights from Any Data

(Level 300 Deck)



SQL Server 2014 and the Data Platform

Mission Critical
Performance



Faster Insights
from Any Data



Platform for
Hybrid Cloud





Faster Insights from Any Data

Easy access to data, big & small

Search, access, and shape internal and external data, and combine with unstructured data

Powerful insights with familiar tools

Speed up analysis and time to insight with the familiarity of Excel in Office and Office 365 and access via mobile devices

Complete BI platform

Scale your BI models and enrich and secure your data with a complete platform spanning cloud and on-premises

Easy Access to Data, Big and Small

Search, access
and shape



Simplify access to public and corporate data

Easily preview, shape, and format your data

Combine with
unstructured



Combine and refine data across multiple sources

Gain insight across relational, unstructured, and semi-structured data

Easily manage
and query



Common management of structured and unstructured data

Query across relational DB and Hadoop with single T-SQL Query

Key features

Project code name “Data Explorer”

Windows Azure Marketplace

Windows Azure HDInsight Service & Microsoft HDInsight Server (preview)

✓ 100% Apache

Parallel Data Warehouse with Polybase

Book1 - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DATA EXPLORER POWERPivot

Online Search

From Web From File From Database From Other Sources

Excel Data Combine Manage Credentials Feedback

Update Settings Fast Combine Options Help

A1

Components - Dow Jones Industrial Average

Dow Jones Industrial Average - Wikipedia, the free encyclopedia

Company	Exchange	Symbol	Industry	Date Added	Notes
3M	NYSE	MMM	Conglomerate	1976-08-09	as Minnesota Mining and
Alcoa	NYSE	AA	Aluminum	1959-06-01	as Aluminum Company of
American Express	NYSE	AXP	Consumer finance	1982-08-30	
AT&T	NYSE	T	Telecommunication	1999-11-01	as SBC Communications
Bank of America	NYSE	BAC	Banking	2008-02-19	
Boeing	NYSE	BA	Aerospace and defense	1987-03-12	
Caterpillar	NYSE	CAT	Construction and mining equipment	1991-05-06	
Chevron Corporation	NYSE	CVX	Oil & gas	2008-02-19	
Cisco Systems	NASDAQ	CSCO	Computer networking	2009-06-08	
Coca-Cola	NYSE	KO	Beverages	1987-03-12	
DuPont	NYSE	DD	Chemical industry	1935-11-20	also 1924-01-22 to 1925-08
ExxonMobil	NYSE	XOM	Oil & gas	1928-10-01	as Standard Oil of New Je
General Electric	NYSE	GE	Conglomerate	1907-11-07	
Hewlett-Packard	NYSE	HPQ	Computers and technology	1997-03-17	
The Home Depot	NYSE	HD	Home improvement retailer	1999-11-01	
Intel	NASDAQ	INTC	Semiconductors	1999-11-01	
IBM	NYSE	IBM	Computers and technology	1979-06-29	
Johnson & Johnson	NYSE	JNJ	Pharmaceuticals	1997-03-17	

Online Search

dji

7 results

Components - Dow Jones Industrial A...
Dow Jones Industrial Average - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Dow_Jones_Ind...

Gold price - Gold as an investment
Gold as an investment - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Gold_as_an_inv...

The statement - The Superinvestors of...
The Superinvestors of Graham-and-Doddsville - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/The_Superinve...

Index levels - United States bear mark...
United States bear market of 2007-2009 - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/United_States_...

READY

Microsoft Power BI for Office 365

Self-service BI with the familiarity of Office and the power of the cloud

Insights in Excel
1 Billion Office Users



Collaborate in Office 365
1 in 4 enterprise customers on Office 365



Discover



Analyze



Visualize



Share



Find



Mobile



Scalable | Manageable | Trusted



Power Query



Power Query



Powerful Self-Service BI with Excel 2013

Discover & Combine



Search and access 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

Power Query – formerly project “Data Explorer”

The screenshot shows the Excel 2013 interface with the 'DATA EXPLORER' ribbon active. The ribbon includes options like 'Online Search', 'From Web', 'From File', 'From Database', 'From Other Sources', 'From Table', 'Merge', 'Append', 'Manage Credentials', 'Send a Smile', 'Send a Frown', 'Update', 'Settings', 'Fast Combine', 'Options', 'Help', and 'About'. The main workspace displays a table titled 'Components - Dow Jones Industrial Average' with columns: Company, Exchange, Symbol, Industry, Date_Added, and Notes. The table lists various companies and their stock information. On the right side, there is an 'Online Search' pane showing search results for 'djia'.

Company	Exchange	Symbol	Industry	Date_Added	Notes
3M	NYSE	MMM	Conglomerate	1976-08-09	as Minnesota Mining and
Alcoa	NYSE	AA	Aluminum	1959-06-01	as Aluminum Company of
American Express	NYSE	AXP	Consumer finance	1982-08-30	
AT&T	NYSE	T	Telecommunication	1999-11-01	as SBC Communications
Bank of America	NYSE	BAC	Banking	2008-02-19	
Boeing	NYSE	BA	Aerospace and defense	1987-03-12	
Caterpillar	NYSE	CAT	Construction and mining equipment	1991-05-06	
Chevron Corporation	NYSE	CVX	Oil & gas	2008-02-19	
Cisco Systems	NASDAQ	CSCO	Computer networking	2009-06-08	
Coca-Cola	NYSE	KO	Beverages	1987-03-12	
DuPont	NYSE	DD	Chemical industry	1935-11-20	also 1924-01-22 to 1925-08
ExxonMobil	NYSE	XOM	Oil & gas	1928-10-01	as Standard Oil of New Je
General Electric	NYSE	GE	Conglomerate	1907-11-07	
Hewlett-Packard	NYSE	HPQ	Computers and technology	1997-03-17	
The Home Depot	NYSE	HD	Home improvement retailer	1999-11-01	
Intel	NASDAQ	INTC	Semiconductors	1999-11-01	
IBM	NYSE	IBM	Computers and technology	1979-06-29	
Johnson & Johnson	NYSE	JNJ	Pharmaceuticals	1997-03-17	

Power Query

Enable self-service data discovery, query, transformation and mashup experiences for Information Workers, via Excel and PowerPivot



Discovery and connectivity to a **wide range of data sources**, spanning volume as well as variety of data.



Highly interactive and intuitive experience for rapidly and iteratively building queries over **any data source, any size**.



Consistency of experience, and parity of query capabilities over all data sources.

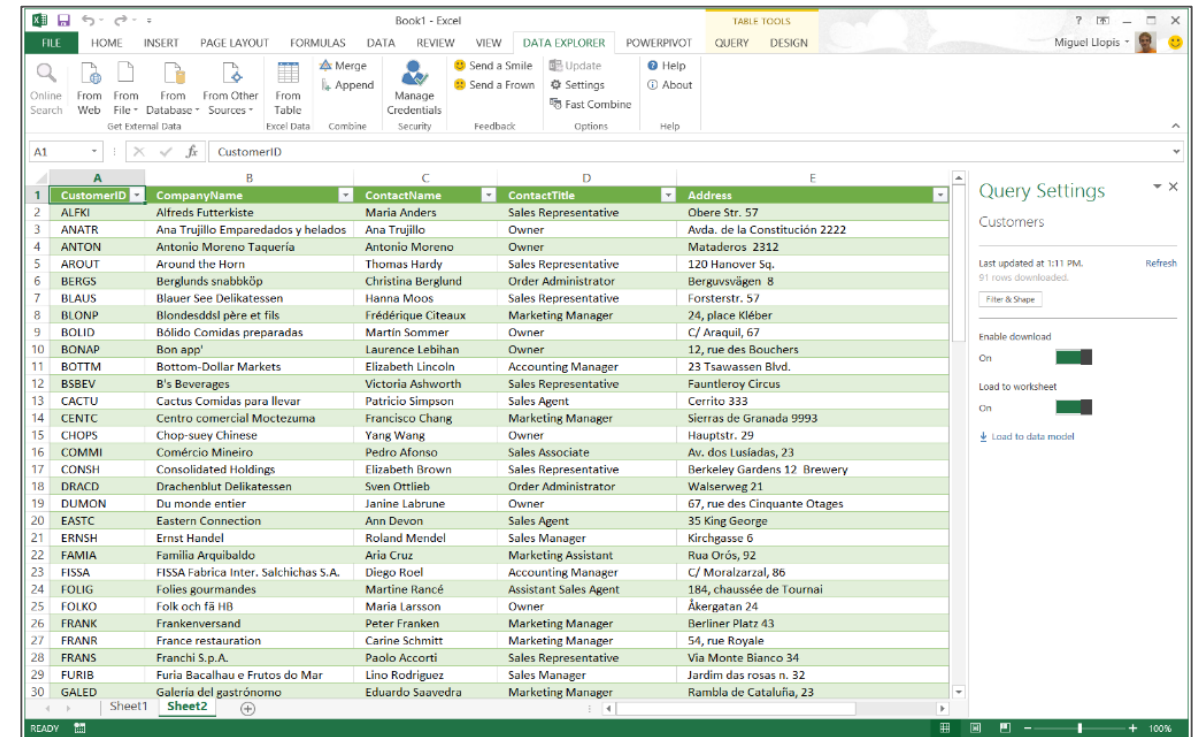


Joins across different data sources; ability to create custom views over data that can then be shared with team/department.

Power Query

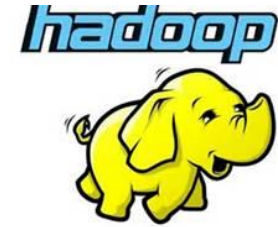
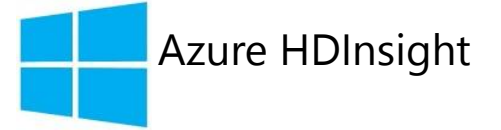
Discover, combine, and refine Big Data, small data, and any data with Data Explorer for Excel.

- 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
ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Obere Str. 57
ANATR	Ana Trujillo Emparedados y helados	Ana Trujillo	Owner	Avda. de la Constitución 2222
ANTON	Antonio Moreno Taquería	Antonio Moreno	Owner	Mataderos 2312
AROUT	Around the Horn	Thomas Hardy	Sales Representative	120 Hanover Sq.
BERGS	Berglunds snabbköp	Christina Berglund	Order Administrator	Berguvsvägen 8
BLAUS	Blauer See Delikatessen	Hanna Moos	Sales Representative	Forsterstr. 57
BLONP	Blondesddsl père et fils	Frédérique Citeaux	Marketing Manager	24, place Kléber
BOLID	Bólido Comidas preparadas	Martin Sommer	Owner	C/ Araquil, 67
BONAP	Bon app'	Laurence Leblan	Owner	12, rue des Bouchers
BOTTM	Bottom-Dollar Markets	Elizabeth Lincoln	Accounting Manager	23 Tsawassen Blvd.
BSBEV	B's Beverages	Victoria Ashworth	Sales Representative	Fauntleroy Circus
CACTU	Cactus Comidas para llevar	Patricio Simpson	Sales Agent	Cerrito 333
CENTC	Centro comercial Moctezuma	Francisco Chang	Marketing Manager	Sierras de Granada 9993
CHOPS	Chop-suey Chinese	Yang Wang	Owner	Hauptstr. 29
COMMI	Comércio Mineiro	Pedro Afonso	Sales Associate	Av. dos Lusladas, 23
CONSH	Consolidated Holdings	Elizabeth Brown	Sales Representative	Berkeley Gardens 12 Brewery
DRACD	Drachenblut Delikatessen	Sven Ottlieb	Order Administrator	Walsenweg 21
DUMON	Du monde entier	Janine Labrune	Owner	67, rue des Cinquante Otages
EASTC	Eastern Connection	Ann Devon	Sales Agent	35 King George
ERNSH	Ernst Handel	Roland Mendel	Sales Manager	Kirchgasse 6
FAMIA	Familia Arquibaldo	Aria Cruz	Marketing Assistant	Rua Orós, 92
FISSA	FISSA Fabrica Inter. Salchichas S.A.	Diego Roel	Accounting Manager	C/ Moralzarzal, 86
FOLIG	Folies gourmandes	Martine Rancé	Assistant Sales Agent	184, chaussée de Tournai
FOLKO	Folk och få Hb	Maria Larsson	Owner	Åkerгатan 24
FRANK	Frankenversand	Peter Franken	Marketing Manager	Berliner Platz 43
FRANR	France restaurant	Carine Schmitt	Marketing Manager	54, rue Royale
FRANS	Franchi S.p.A.	Paolo Accorti	Sales Representative	Via Monte Bianco 34
FURIB	Furia Bacalhau e Frutos do Mar	Uino Rodriguez	Sales Manager	Jardim das rosas n. 32
GALED	Galería del gastrónomo	Eduardo Sauvedra	Marketing Manager	Rambla de Catalunya, 23

Data Sources





PowerPivot



PowerPivot



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

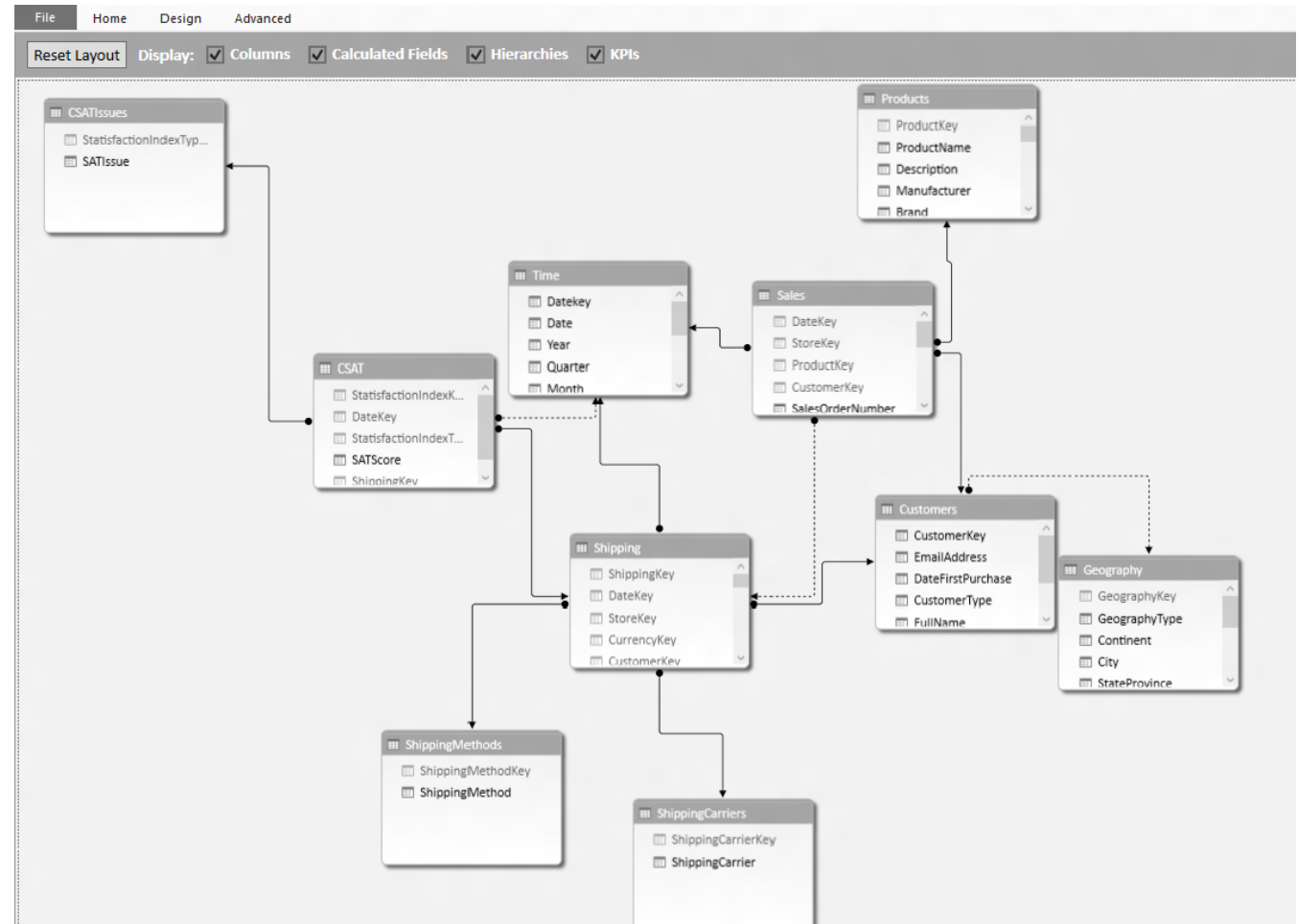
Visualize & Explore



Bring your data to life with interactive visualization

Explore data in new ways to discover hidden insights

PowerPivot



Introducing PowerPivot



- PowerPivot empowers business users to create self-service BI data models in Excel
 - Achieved with a client-side version of SQL Server Analysis Services, known as the In-Memory Analytics Engine
 - Can efficiently store data volumes far greater than what Excel worksheets can achieve
- A separate window can be used to load, explore, relate, and enrich data with calculations
- Can import and relate data from corporate, local, and ad hoc data stores
- In Excel 2013 Professional Plus edition, PowerPivot is installed, but not enabled

PowerPivot for SharePoint

- PowerPivot for SharePoint extends SharePoint 2013 and Excel Services to add server-side processing, collaboration, and document management support for the PowerPivot workbooks published to SharePoint
- The PowerPivot Gallery, a special type of document library, is provided to browse published PowerPivot workbooks and to configure automatic data refresh
- The PowerPivot Service deploys the embedded data model to an Analysis Service instance within the SharePoint farm, and Excel Services is used to query the deployed data model
- Data connections can be defined to query the PowerPivot data model by using the document URL



Power View



Power View



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 xVelocity in-memory technology

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

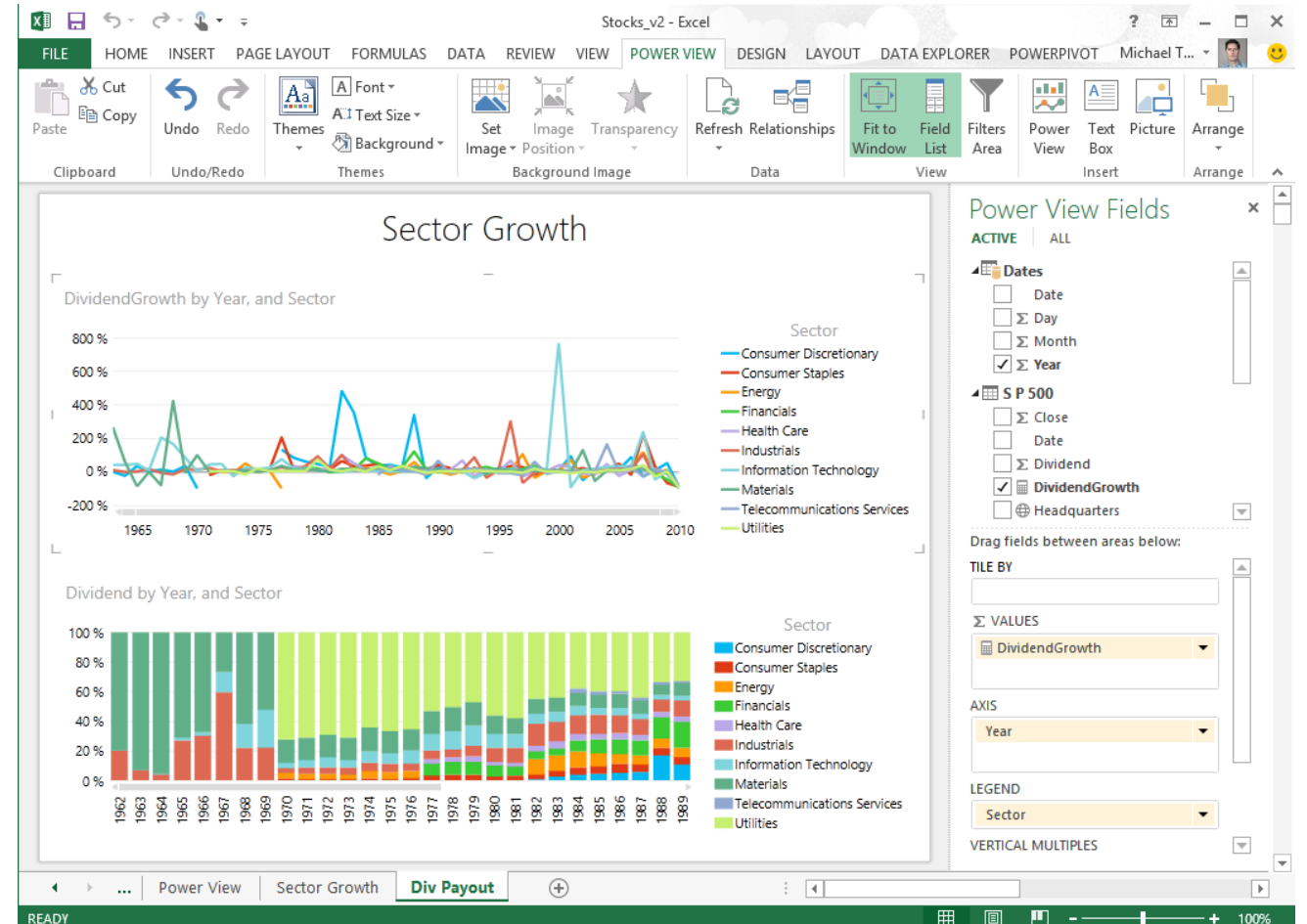
Visualize & Explore



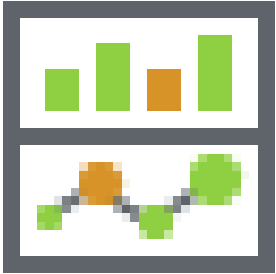
Bring your data to life with interactive visualization

Explore data in new ways to uncover hidden insights

Power View



Introducing Power View

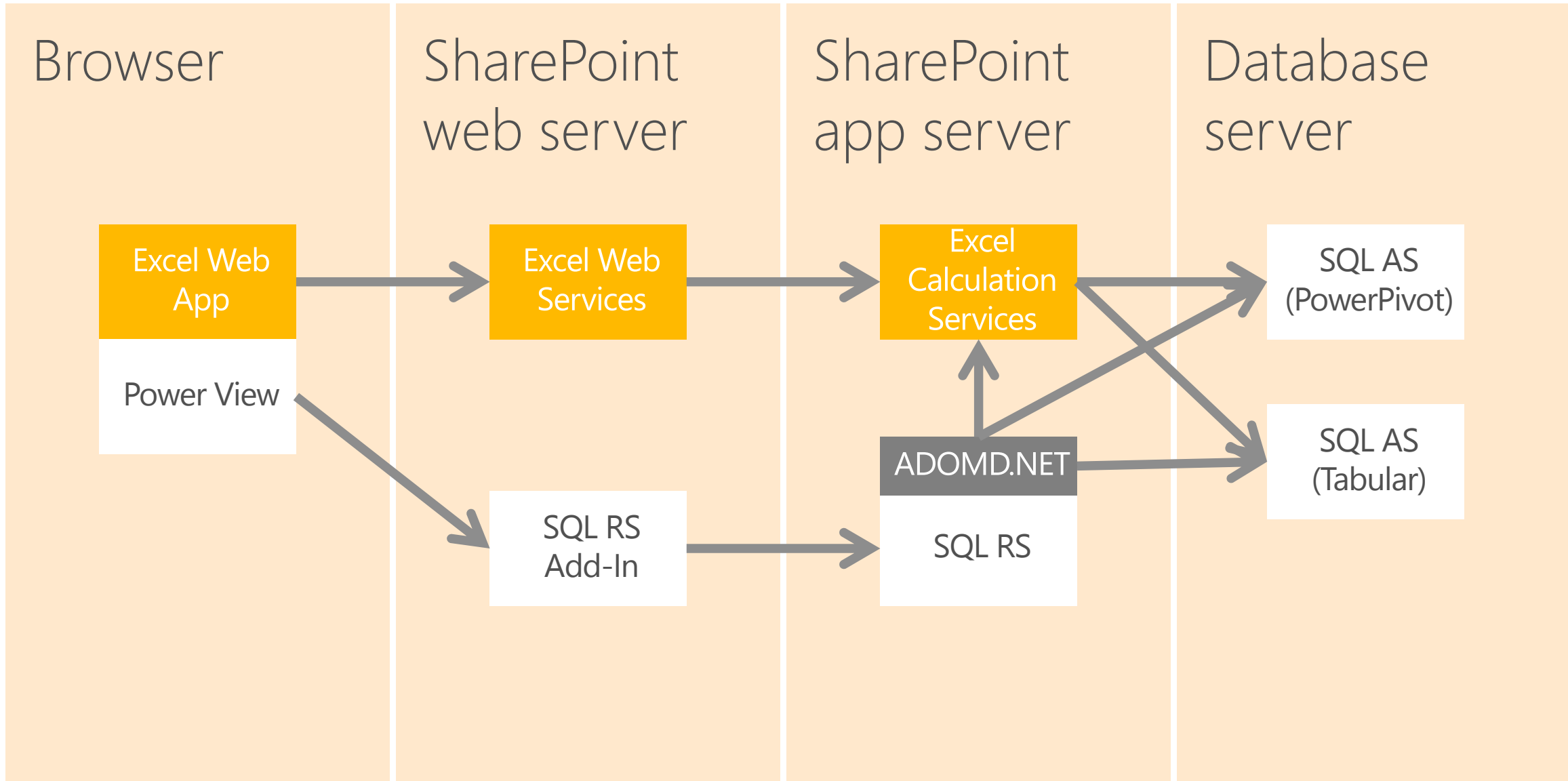


- Power View is an interactive data exploration, visualization, and presentation experience
 - Highly visual design experience
 - Rich meta-driven interactivity
 - Presentation-ready at all times
- It delivers intuitive ad-hoc reporting for business users
- Reports can be based on tabular data models, including PowerPivot data models
- It is now also available in Excel 2013, and with new features:
 - Maps
 - Pie charts
 - Hierarchies
 - KPIs
 - Drill down/Drill up
 - Report styles, themes and text resizing
 - Backgrounds with images
 - Hyperlinks
 - Printing

Power View in Excel



Power View in SharePoint

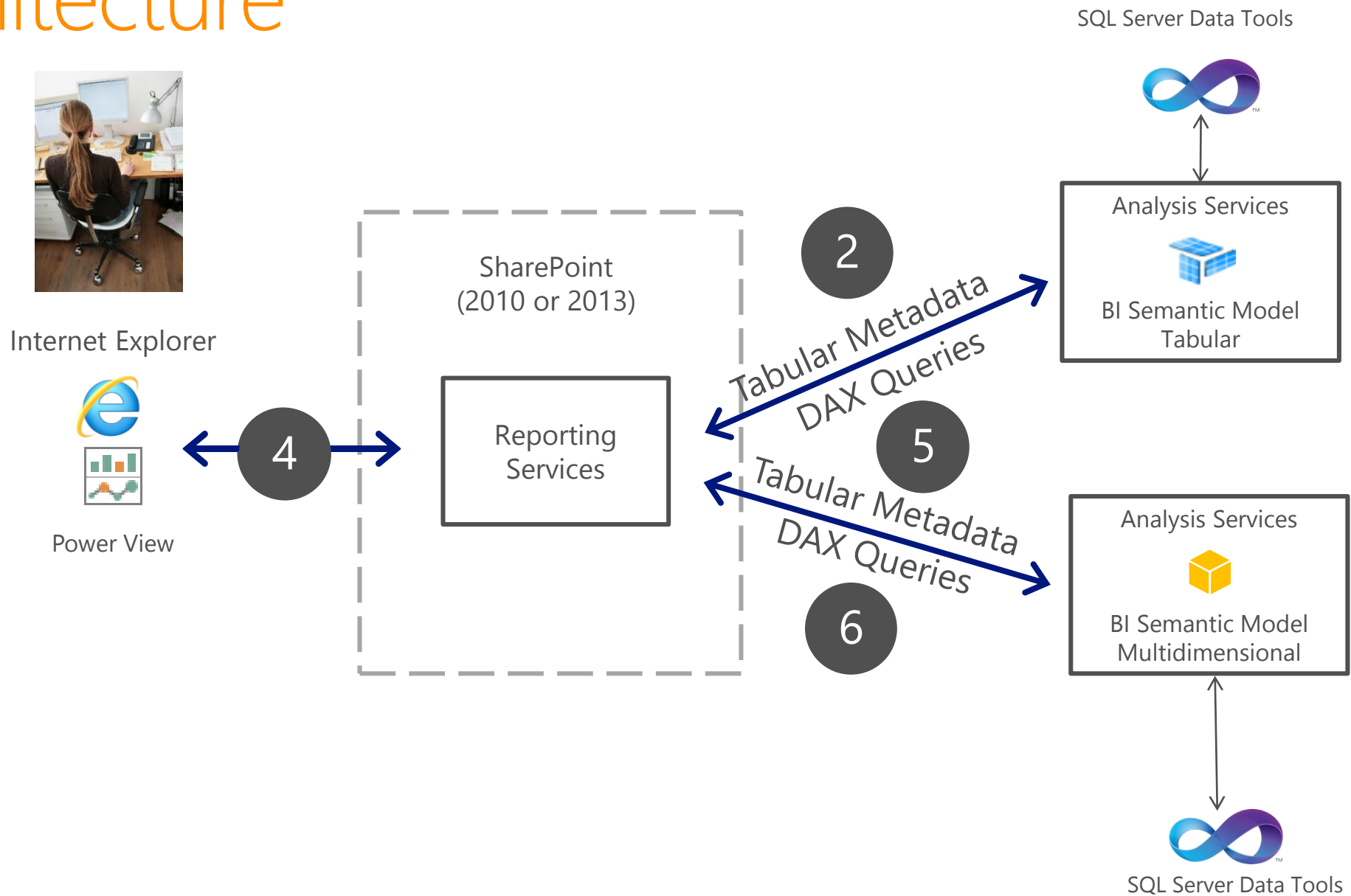


Power View for Multidimensional Models

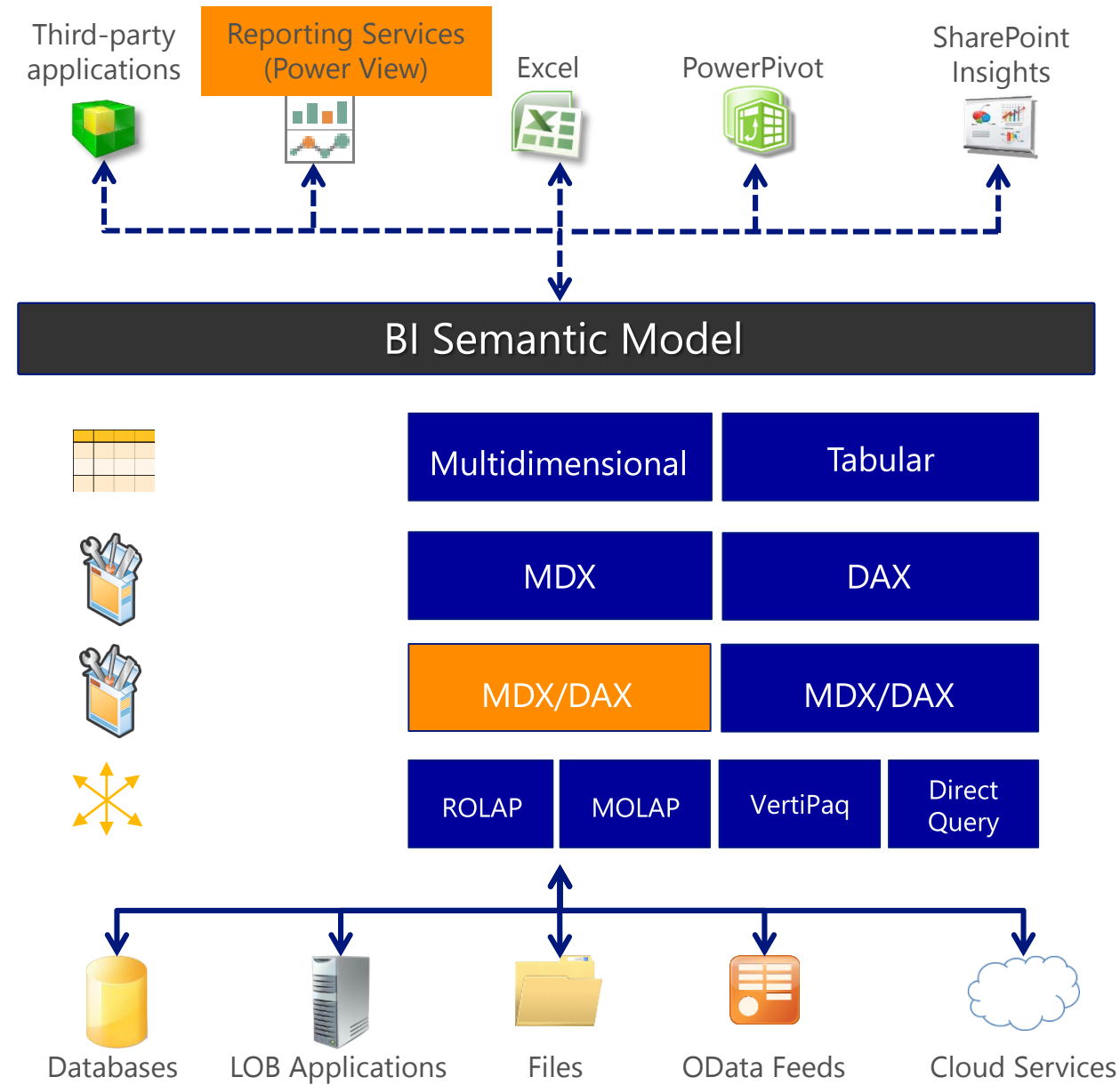
- Power View on Analysis Services via BISM
- Native support for DAX in Analysis Services
- Better flexibility: Choice of DAX on Tabular or Multidimensional (cubes)



Architecture



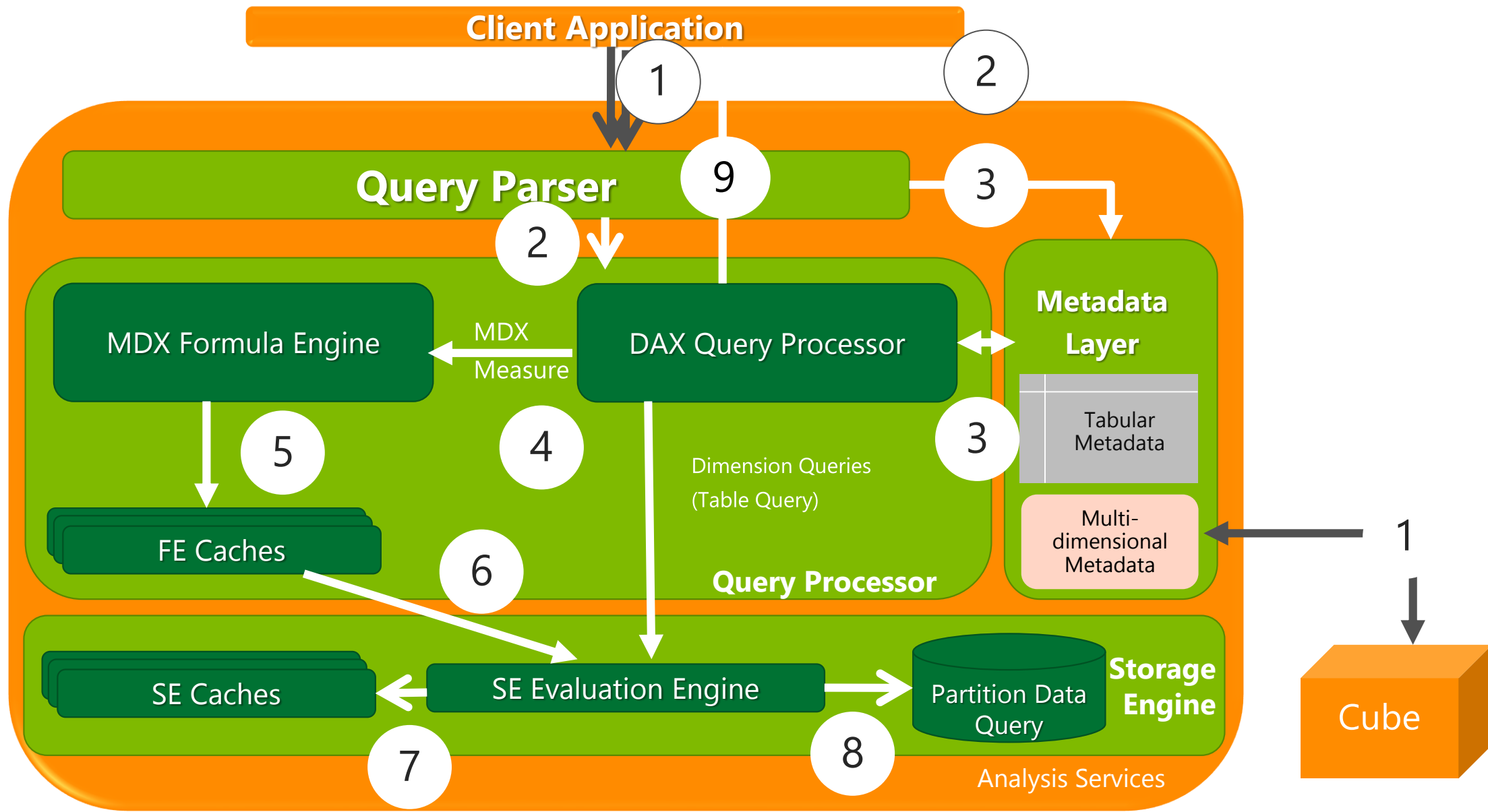
BI Semantic Model: Architecture



Multidimensional-Tabular Mapping

BISM-MD Object	Tabular Object
Cube	Model
Cube Dimension	Table
Attributes (Key(s), Name)	Columns
Measure Group	Table
Measure	Measure
Measure without MeasureGroup	Within Table called "Measures"
Measuregroup→Cube Dimension relationship	Relationship
Perspective	Perspective
KPI	KPI
User/Parent-Child Hierarchies	Hierarchies

Query Execution Architecture





Power Map



Power Map



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 xVelocity in-memory technology

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

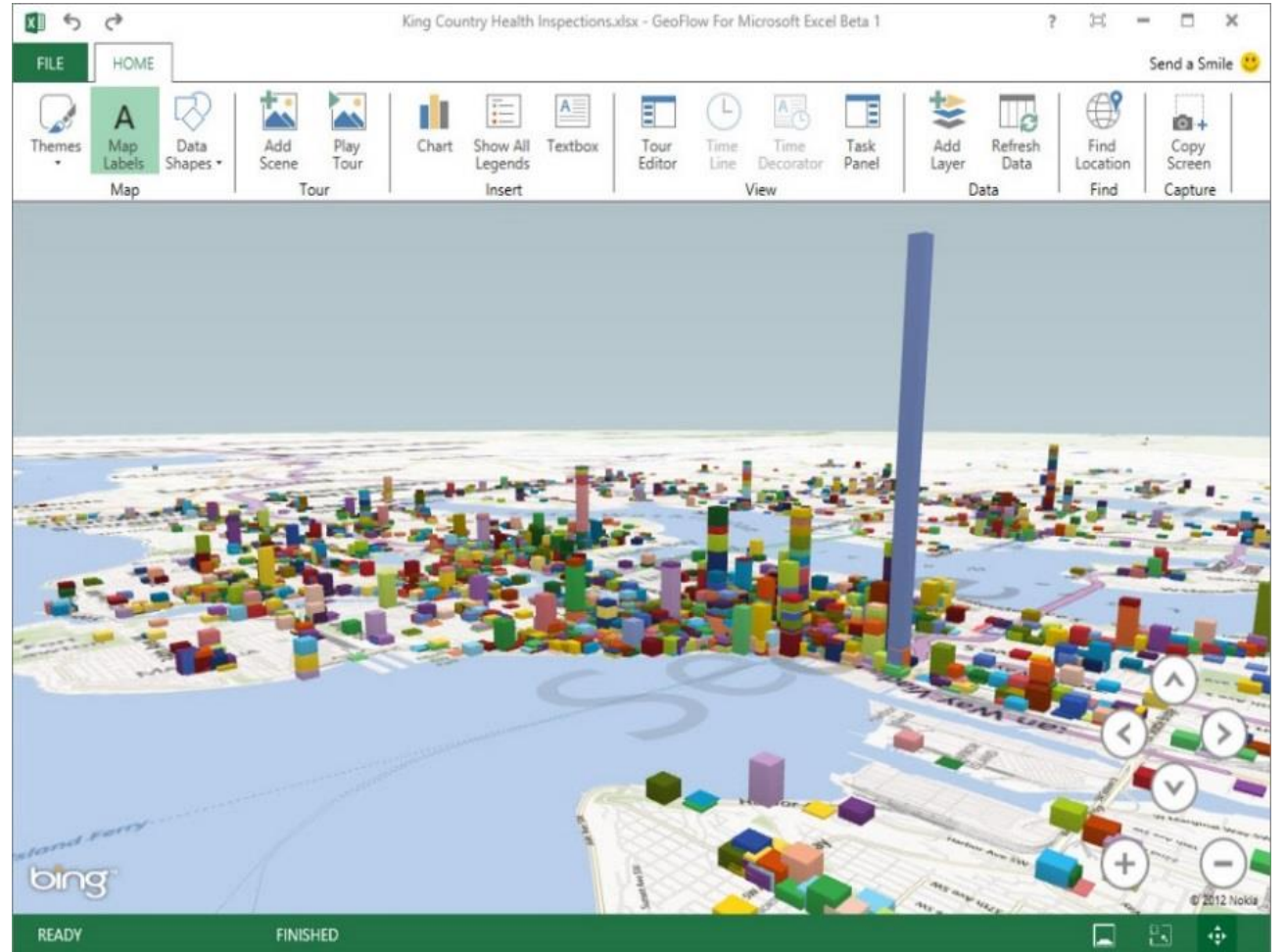
Visualize & Explore



Bring your data to life with interactive visualization

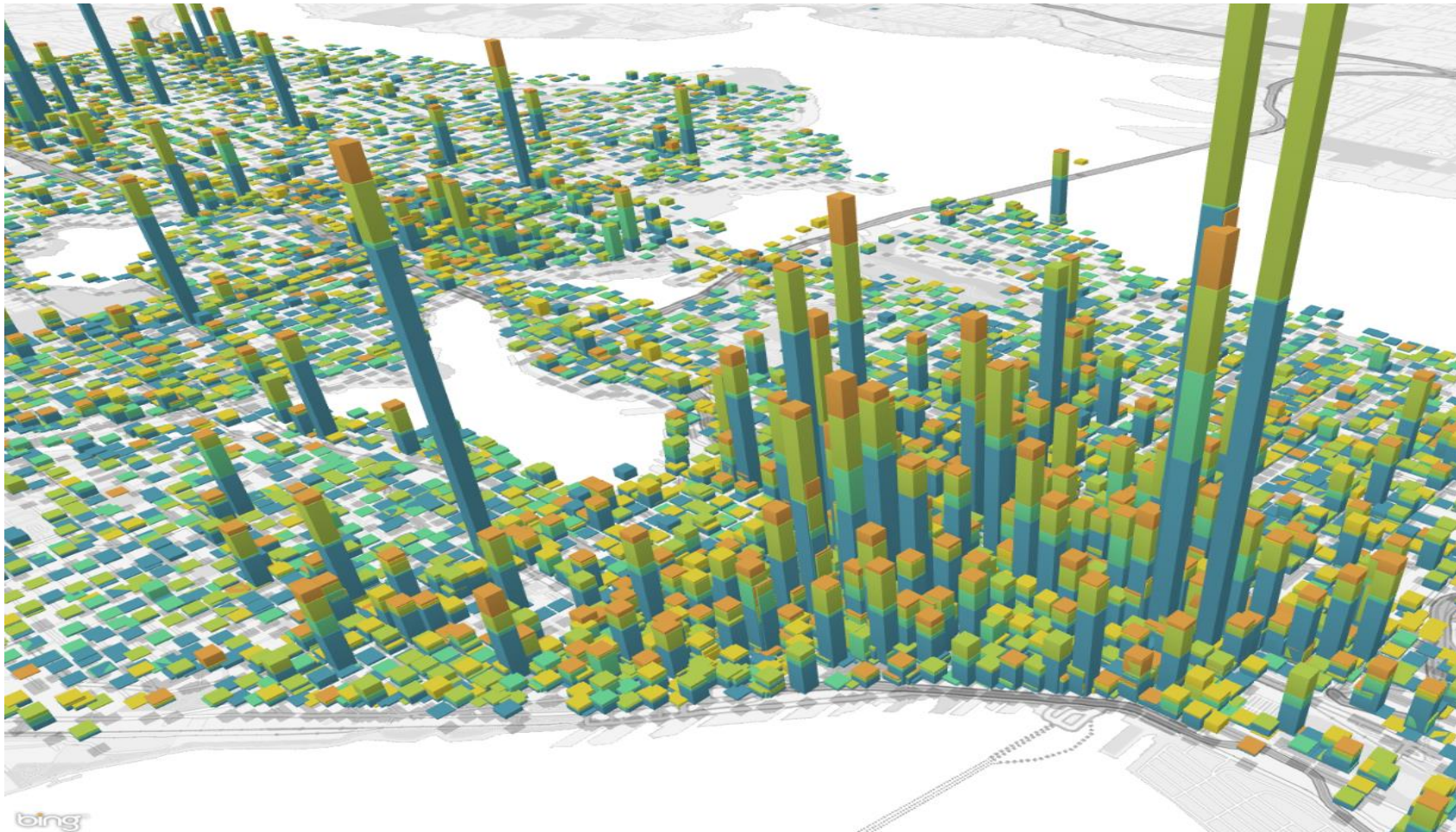
Explore data in new ways to discover hidden insights

Power Map – formerly project “Geoflow”

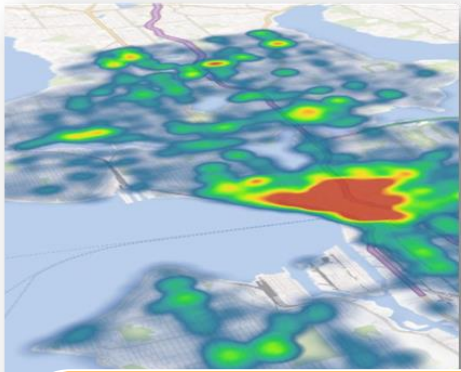


What Is Power Map?

Power Map for Microsoft Excel enables information workers to discover and share new insights from geographical and temporal data through three-dimensional storytelling.

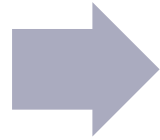


Power Map: Steps to 3D insights



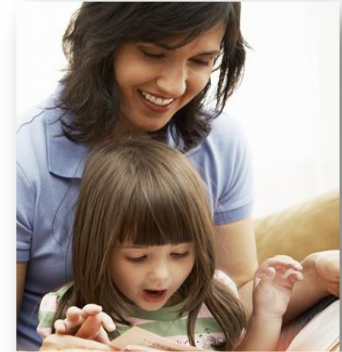
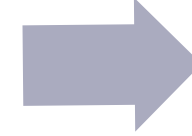
Map Data

- Data in Excel
- Geo-Code
- 3D and 3 Visuals



Discover Insights

- Play over Time
- Annotate points
- Capture scenes

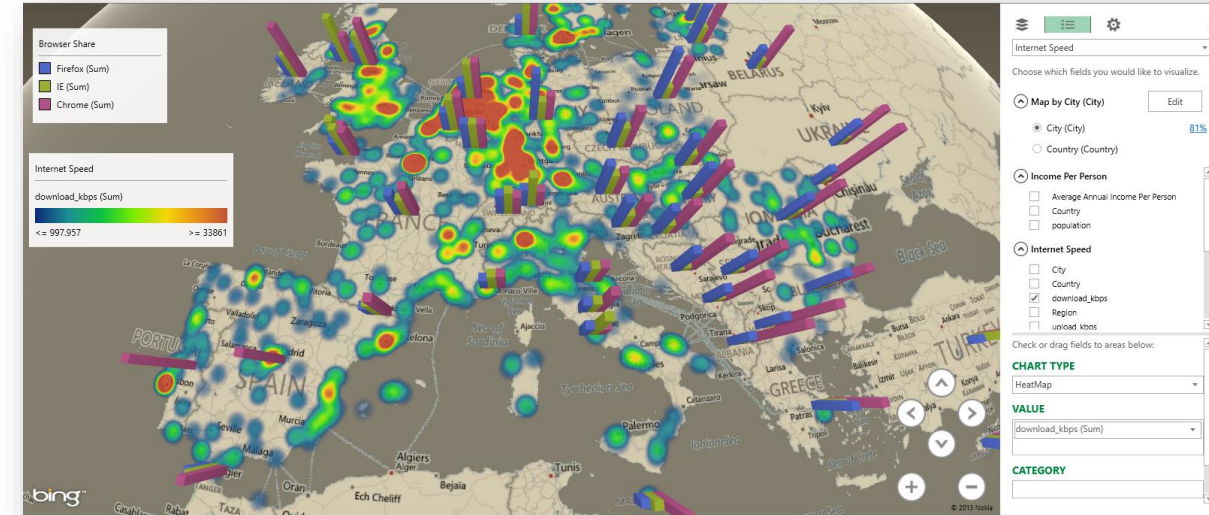
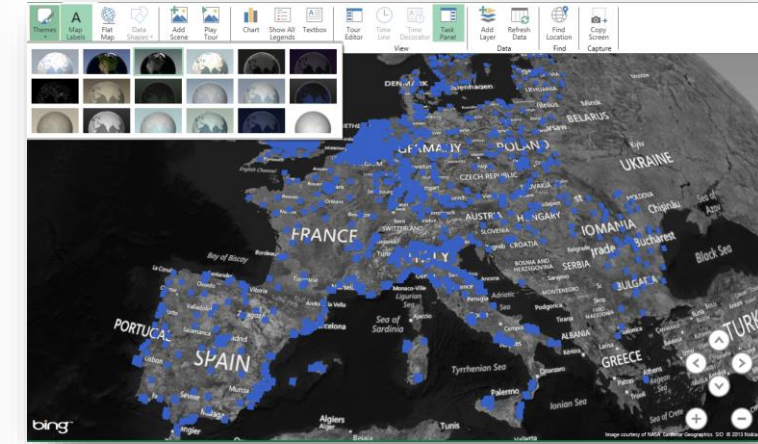
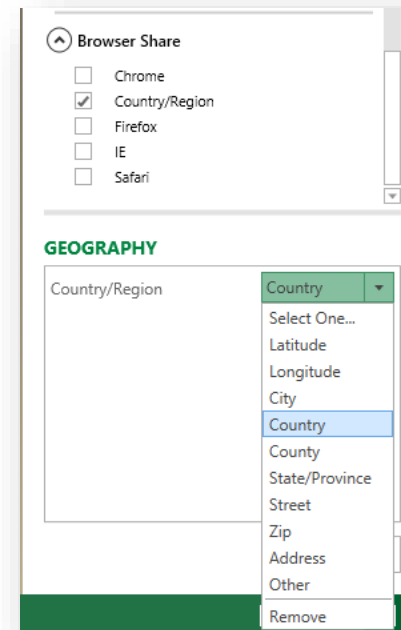


Share Stories

- Cinematic Effects
- Interactive Tours
- Share Workbook

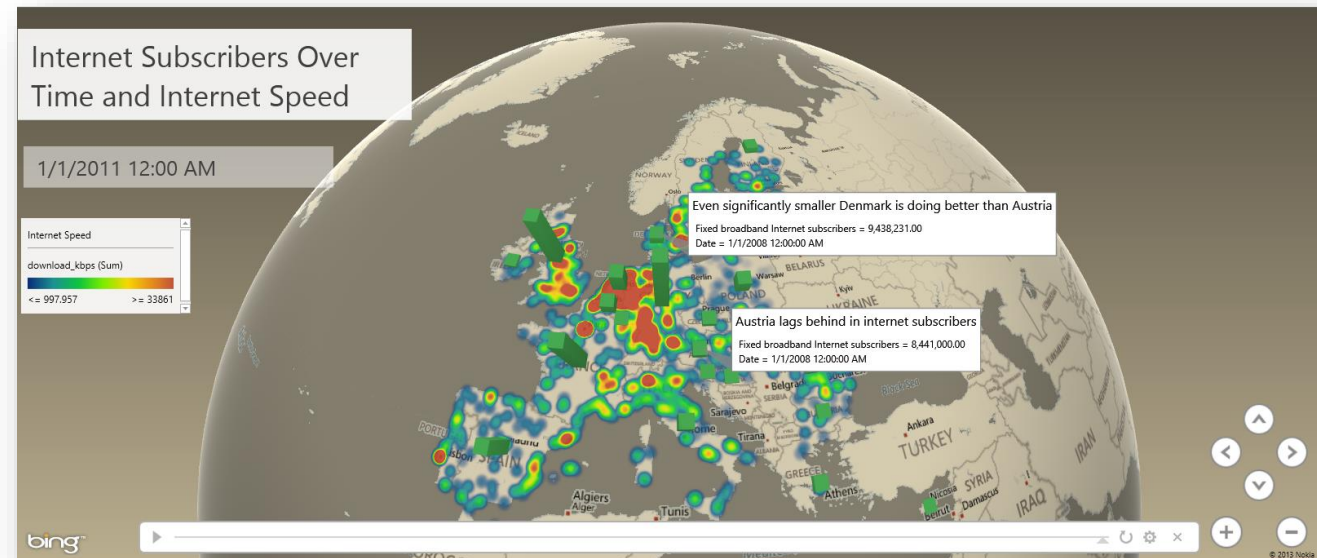
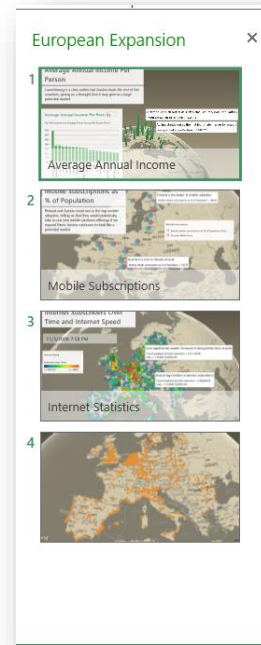
Map Data

- 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



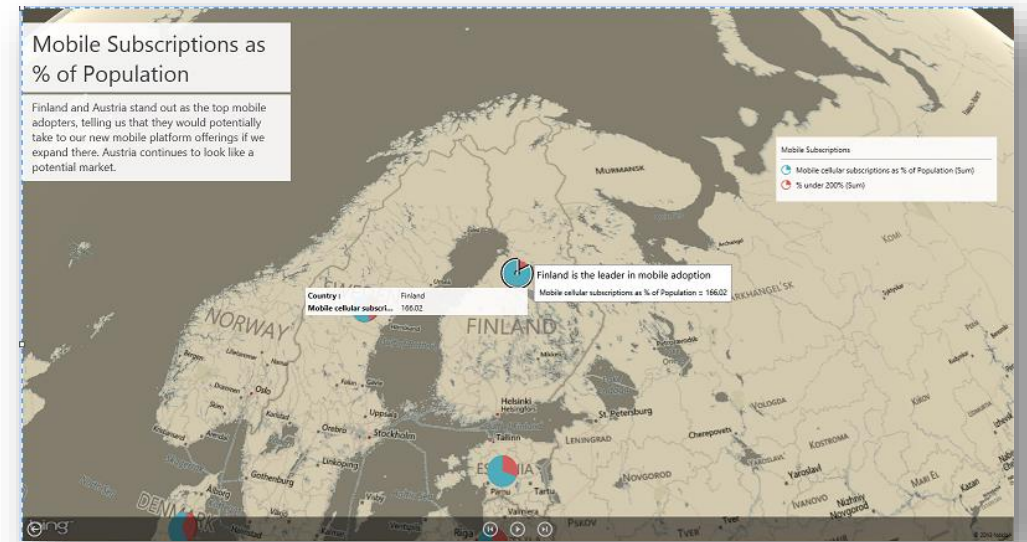
Discover Insights

- Play your data over time
- Make direct comparisons in 2D charts
 - Flat 2D map for comparison of 3D columns
- Annotate insights with metadata
 - Can add images to annotations.
- Capture "Scenes" like slides in a show



Share Stories

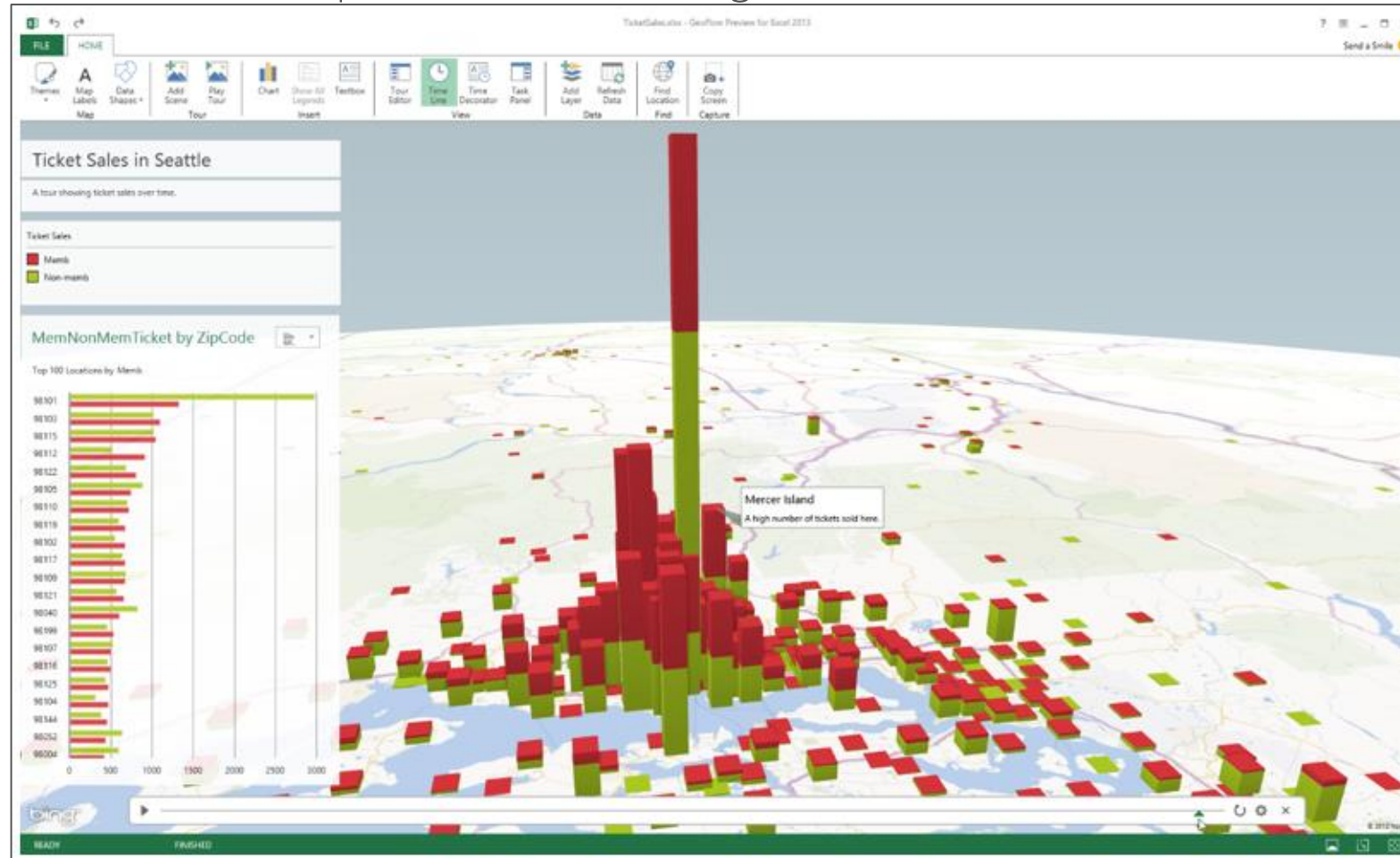
- Weave scenes with cinematic effects
- Interact and resume during presentation
- Share broadly within the workbook
 - Export to Video for Viral!



Power Map

Excel Add-in to Enhance Data Visualization

Map data, discover insight, and share stories





Power BI Site



Power BI Site



Collaborate and Stay Connected with Office 365

Share Data & Insights



Easily setup powerful BI sites

Share live reports with data refresh from on-prem sources

Manage and share data sets and monitor usage

Ask Questions & Get Answers

Ask questions in the natural language speech bubble

Get back answers as the system generates data visualizations

Q&A

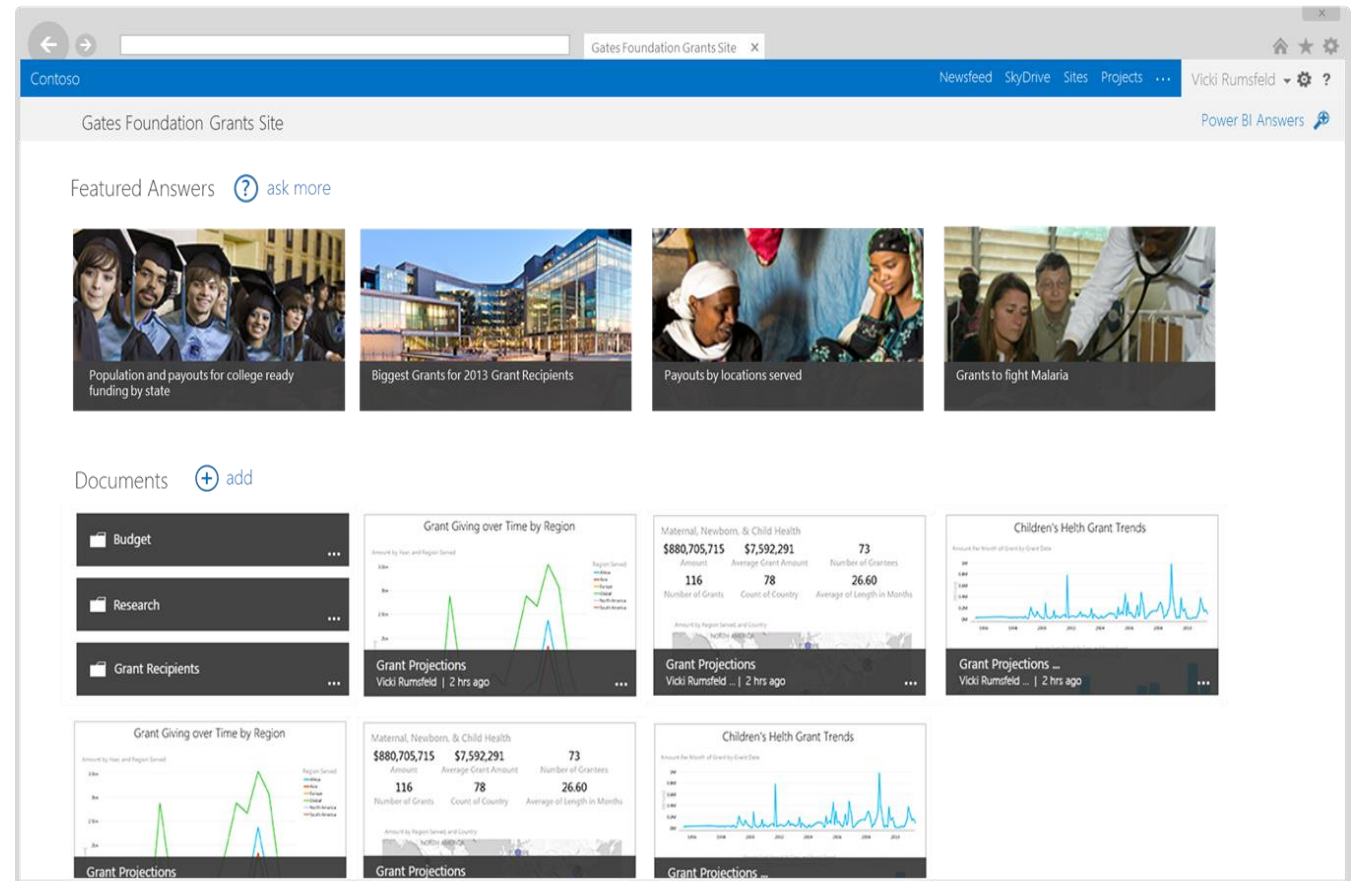
Connect from Anywhere



View Power View reports on any device with HTML5 support

Discover and explore all your favorite reports with the mobile BI app

Power BI Site in Office 365





Q & A



Q & A



Collaborate and Stay Connected with Office 365

Share Data & Insights



Easily setup powerful BI sites

Share live reports with data refresh from on-prem sources

Manage and share data sets and monitor usage

Ask Questions & Get Answers

Ask questions in the natural language speech bubble

Get back answers as the system generates data visualizations

Q&A

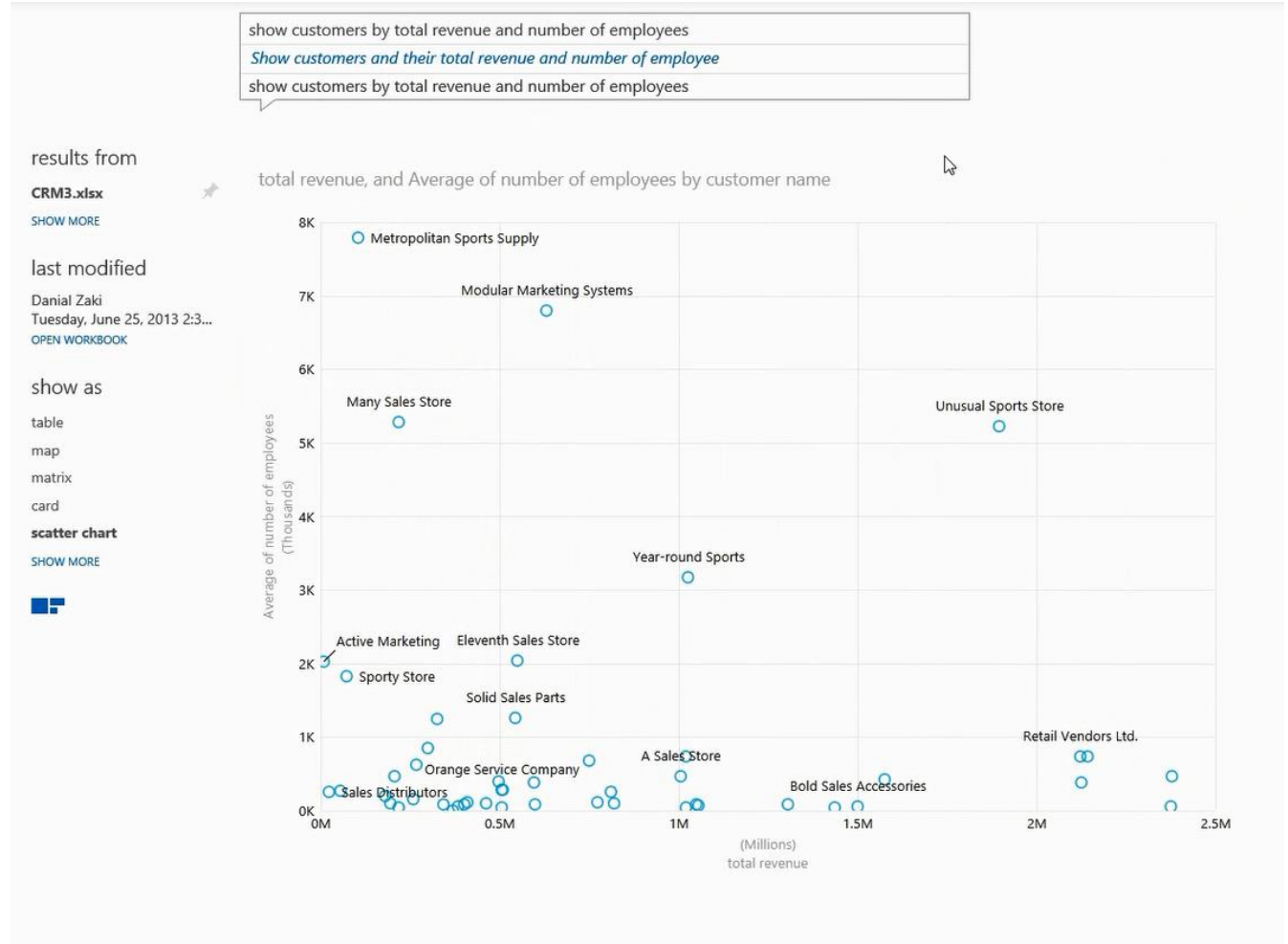
Connect from Anywhere



View Power View reports on any device with HTML5 support

Discover and explore all your favorite reports with the mobile BI app

Data Stewardship experience





Mobile BI



Mobile BI



Collaborate and Stay Connected with Office 365

Share Data & Insights



Easily setup powerful BI sites

Share live reports with data refresh from on-prem sources

Manage and share data sets and monitor usage

Ask Questions & Get Answers

Ask questions in the natural language speech bubble

Get back answers as the system generates data visualizations

Q&A

Connect from Anywhere



View Power View reports on any device with HTML5 support

Discover and explore all your favorite reports with the mobile BI app

HTML5 & Mobile BI app



Enable Deep Business and Customer Connections

Virtually Anytime, Anywhere



Boost agility with real-time access to apps and data from anywhere



Engage customers with smart, contextual mobile experiences

Excite and Engage Users

Deliver Immersive, Connected Customer Experiences

Beautiful
experiences +
security and
performance



Connect with
social apps and
networks



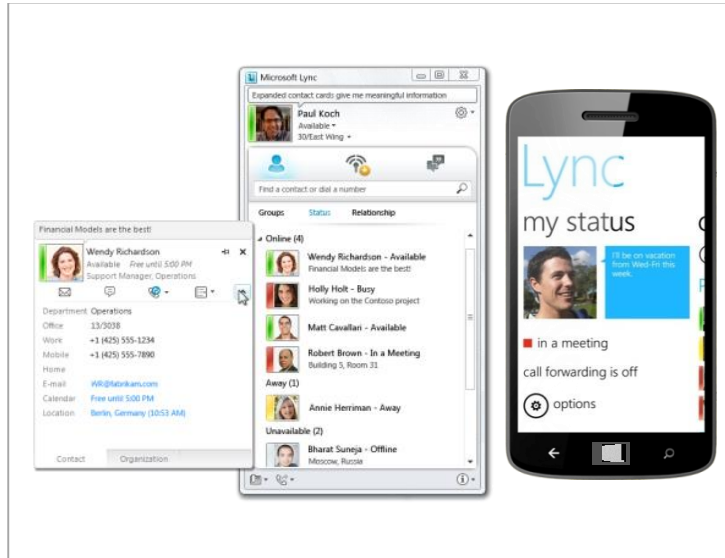
Real time content
and updates



Optimize for
discovery and
reach

Stay Productive on the Go

Deliver Familiar, Connected Experiences to a Mobile Workforce



Integrate your communications



Work together in real time



Simplify mobile access to business information

...while ensuring enterprise security, manageability, and compliance

Mobile BI Capabilities Available Today

Browser-based corporate BI solutions on iOS, Android and Windows:

- SharePoint Mobile enhancements
- PerformancePoint Services
- Excel Services
- SQL Server Reporting Services

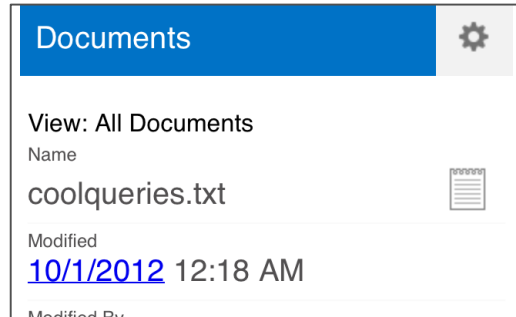


"Ultimately, the new Microsoft mobile BI solution leads to more revenue for Recall and gives us deeper customer insight, helping us stay ahead of our competitors."

Recall Records Management Company Gets Real-Time BI, Boosts Sales with Mobile Solution case study. [Full Case study](#).

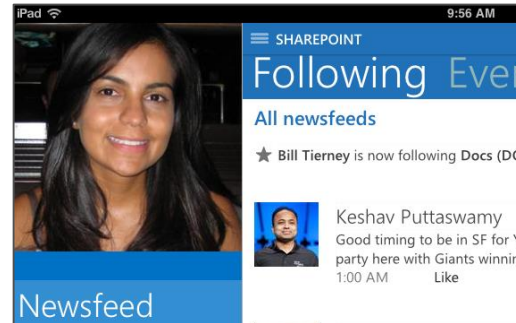
SharePoint Mobile 2013

Mobile Browser



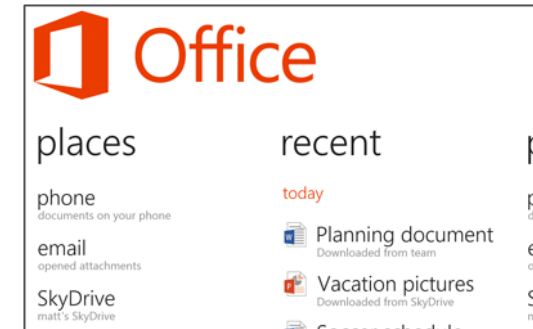
Support across different mobile devices, including touch for tablets and phone

Native Apps



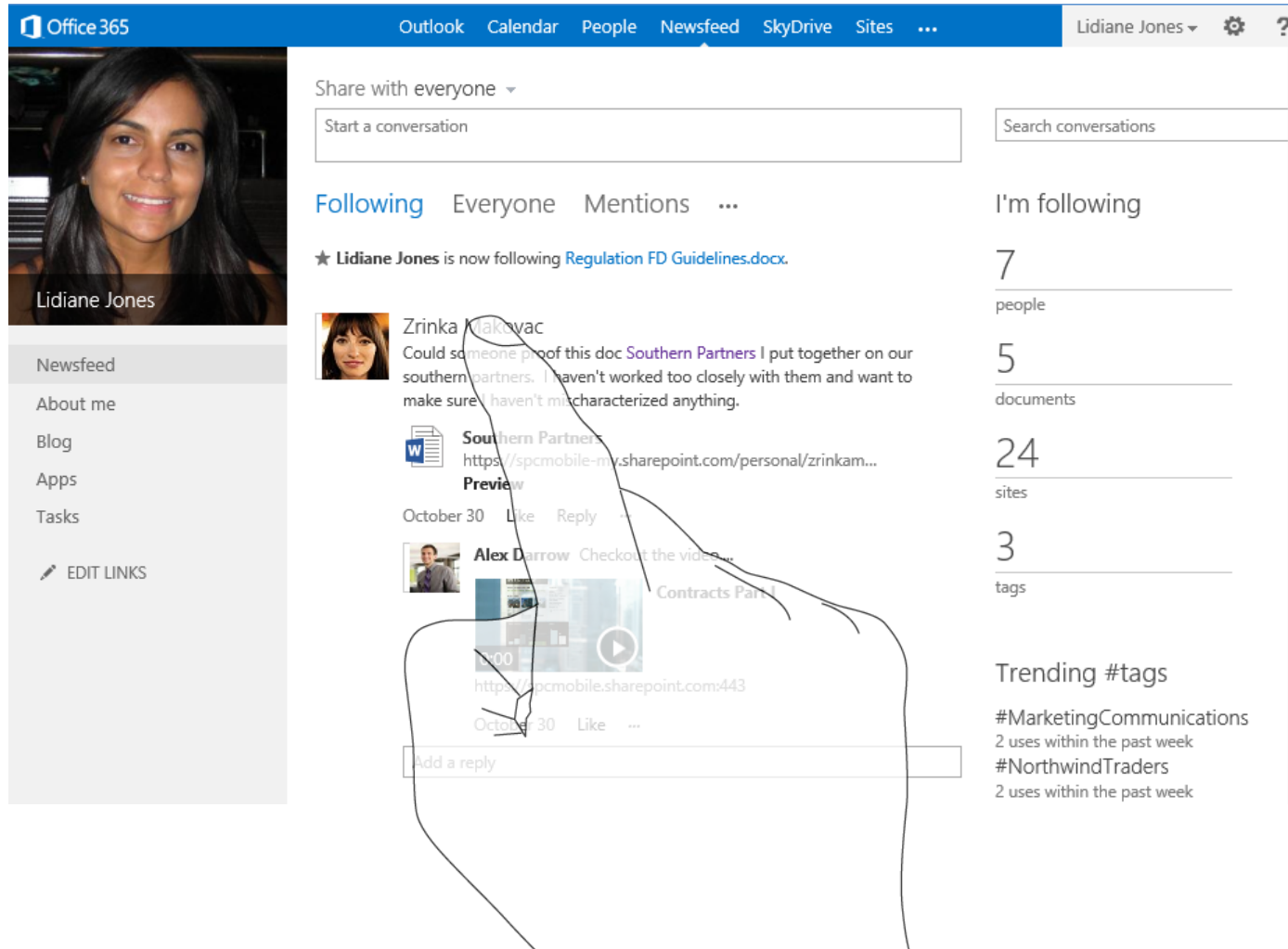
Rich native apps experience to business social interactions and collaboration

Office Hub



A hub to all your doc storage services and Office rich editing experience

Tablet Touch Experience

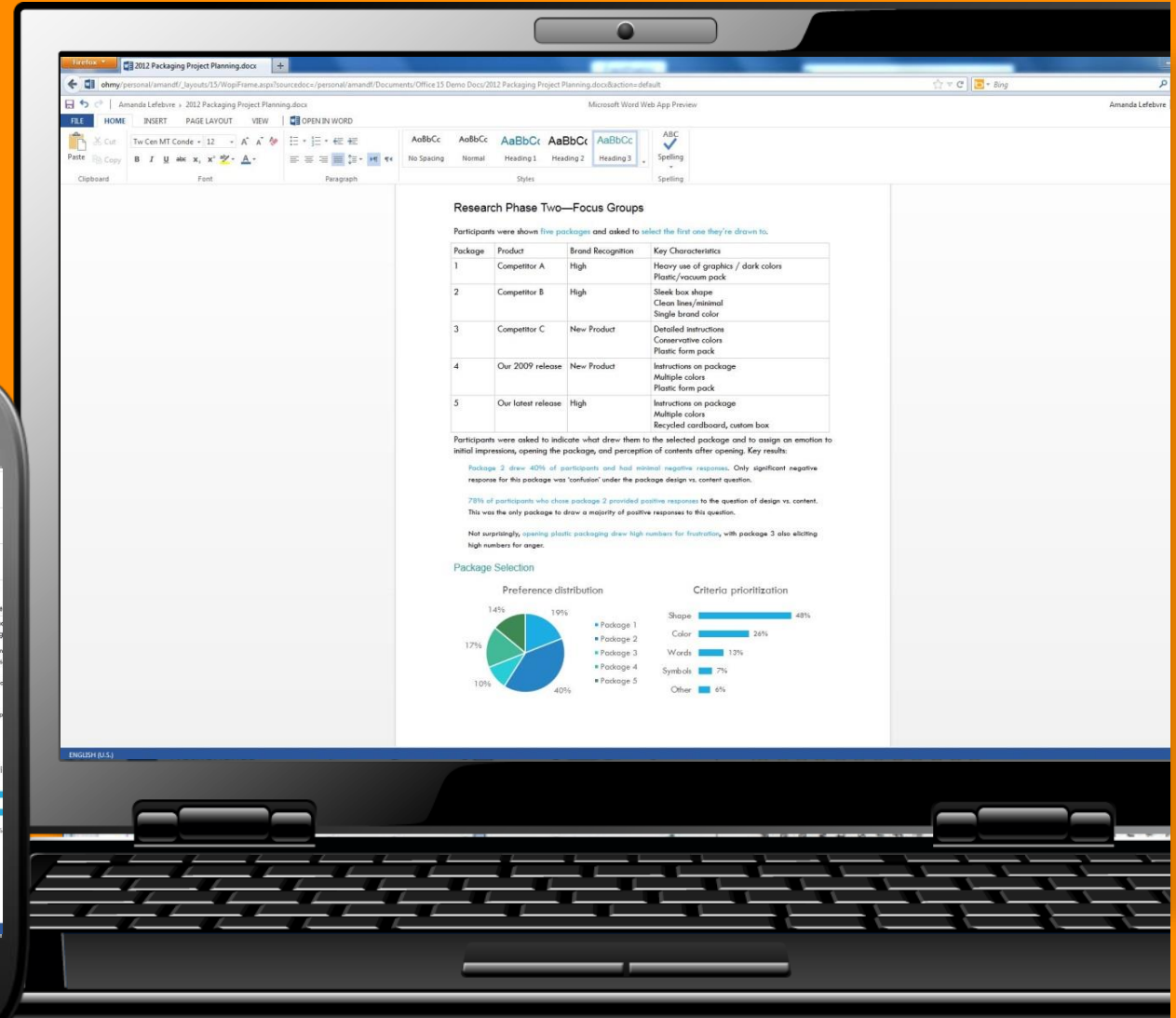
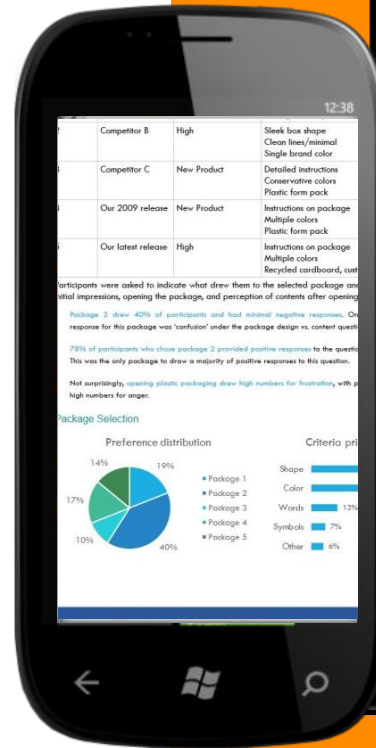


- Clean user experience
- Large touch targets
- Filtering and navigation is as simple as working on the desktop

Office Web Apps

Never be without
the tools you need.

Access and share with
confidence.



Excel Web App

Excel reports, dashboards, pivot tables, charts in a browser

- Gallery View: optimized for touch

Excel Web App allows lightweight editing on Excel reports

- High-fidelity viewing, interaction with advanced analysis views
- External connections and refresh connections
- Simultaneous coauthoring
- Embedding and the Excel button

Excel Web App

Mobile support for Windows, iOS, Android, and other mobile devices

Quick Explore

- Enhanced visual, browser-based data exploration
- Drill in and out
- Pivot Tables and Pivot Charts



Mobile-Friendly Apps for Office

Office Apps can be developed using standard web technologies such as HTML, XML, CSS, JavaScript, REST APIs

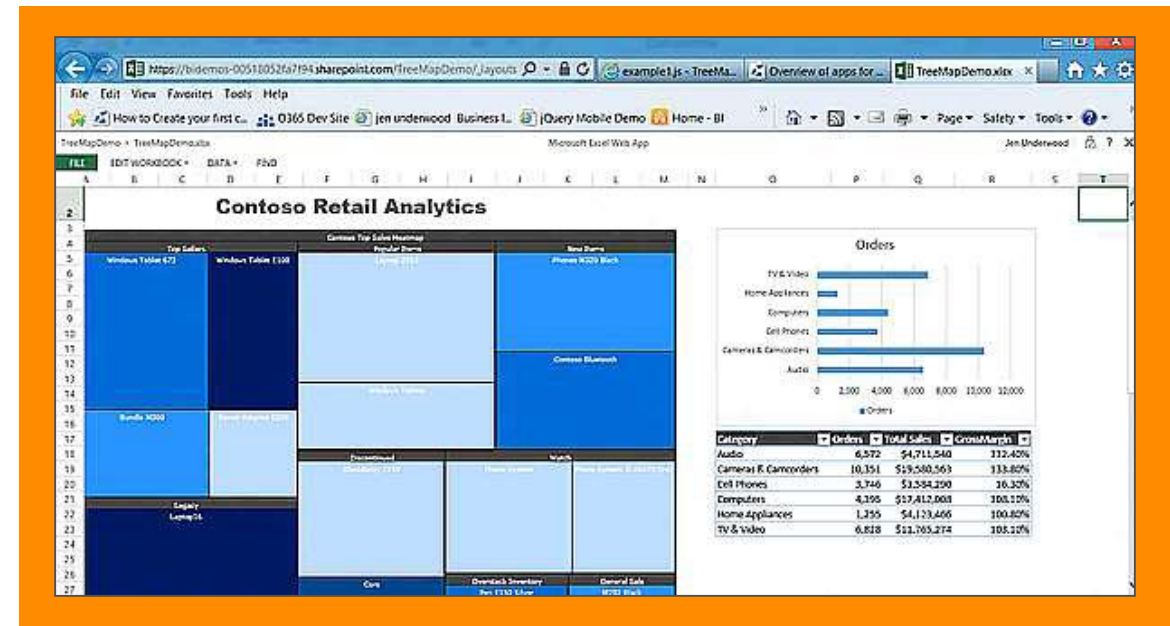
Office App APIs can read/write data, interact with Office client

- Office desktop clients and O365 Office Web Apps
- Mobile device, touch friendly

Get creative with open JavaScript libraries like D3 or InfoVis Toolikt

Get started with Office 365

Developer account



Extend with Hybrid Cloud Solutions

On-Premise Data Access



Enable manual or scheduled data refresh for reports in Office 365 to on-premise sources

Enterprise Data Search



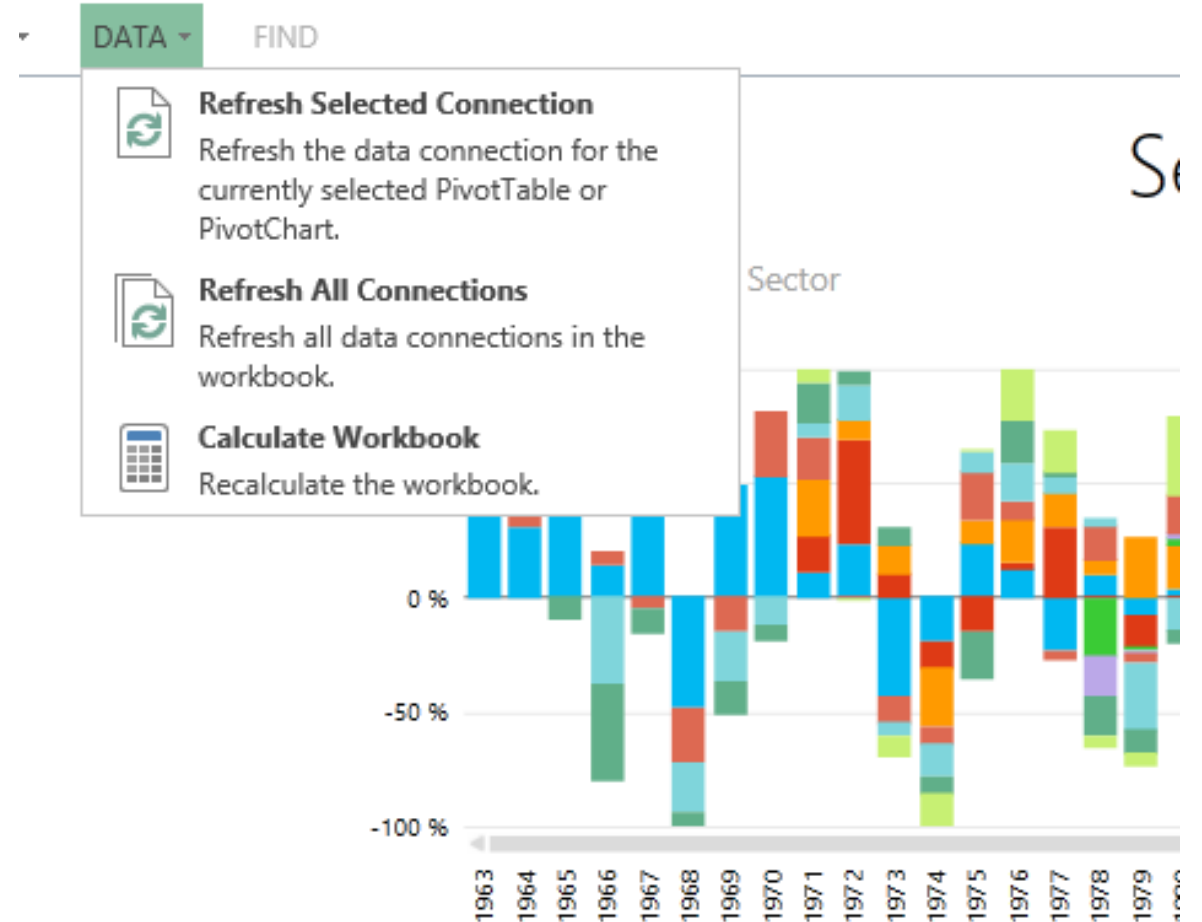
Enable ease of discovery with data search across on premise data

Data Usage & Telemetry



Monitor data usage across your organization to optimize system investments

Data Management Gateway



Extend with Hybrid Cloud Solutions

On-Premise Data Access



Enable manual or scheduled data refresh for reports in Office 365 to on-premise sources

Enterprise Data Search



Enable ease of discovery with data search across on premise data

Data Usage & Telemetry



Monitor data usage across your organization to optimize system investments

Data Catalogue

The screenshot displays the Microsoft Excel interface with the 'Data Explorer' tab selected in the ribbon. The 'Online Search' pane is open on the right side, showing search results for 'download metrics for new products'. The results list includes items like 'DataExplorer StandAlone Com...', 'DataExplorer StandAlone Atte...', 'Production - Cereal', '2006 worldwide pork consum...', 'Real GDP Per Capita by State (...)', and 'Current Dollars GDP All Indust...'. The search results are displayed in a list format with source information and a brief description. The Excel spreadsheet area is visible in the background, showing a grid with columns A through L and rows 1 through 29. The status bar at the bottom indicates 'Sheet1' and 'READY'.

Extend with Hybrid Cloud Solutions

On-Premise Data Access



Enable manual or scheduled data refresh for reports in Office 365 to on-premise sources

Enterprise Data Search



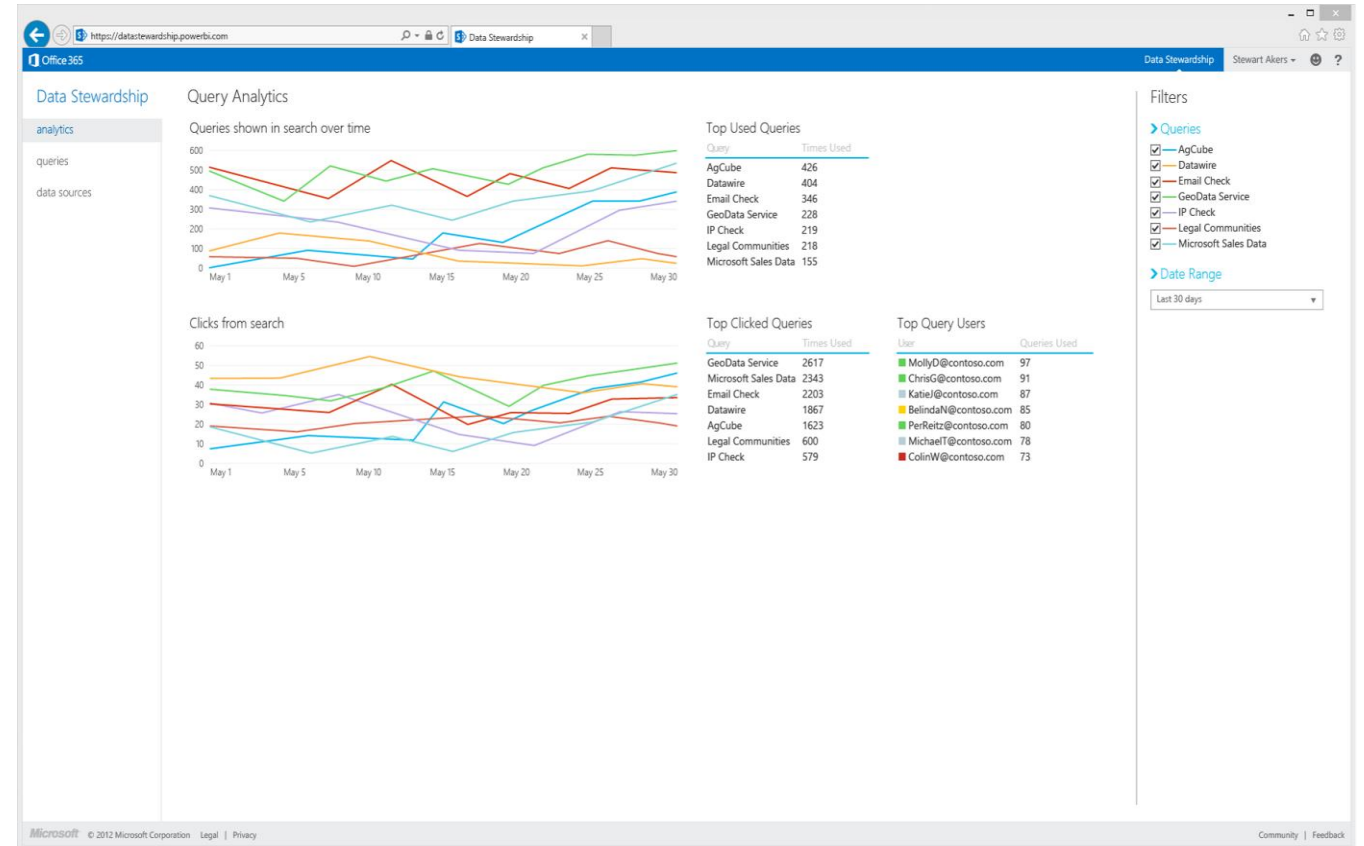
Enable ease of discovery with data search across on premise data

Data Usage & Telemetry



Monitor data usage across your organization to optimize system investments

Data Catalogue





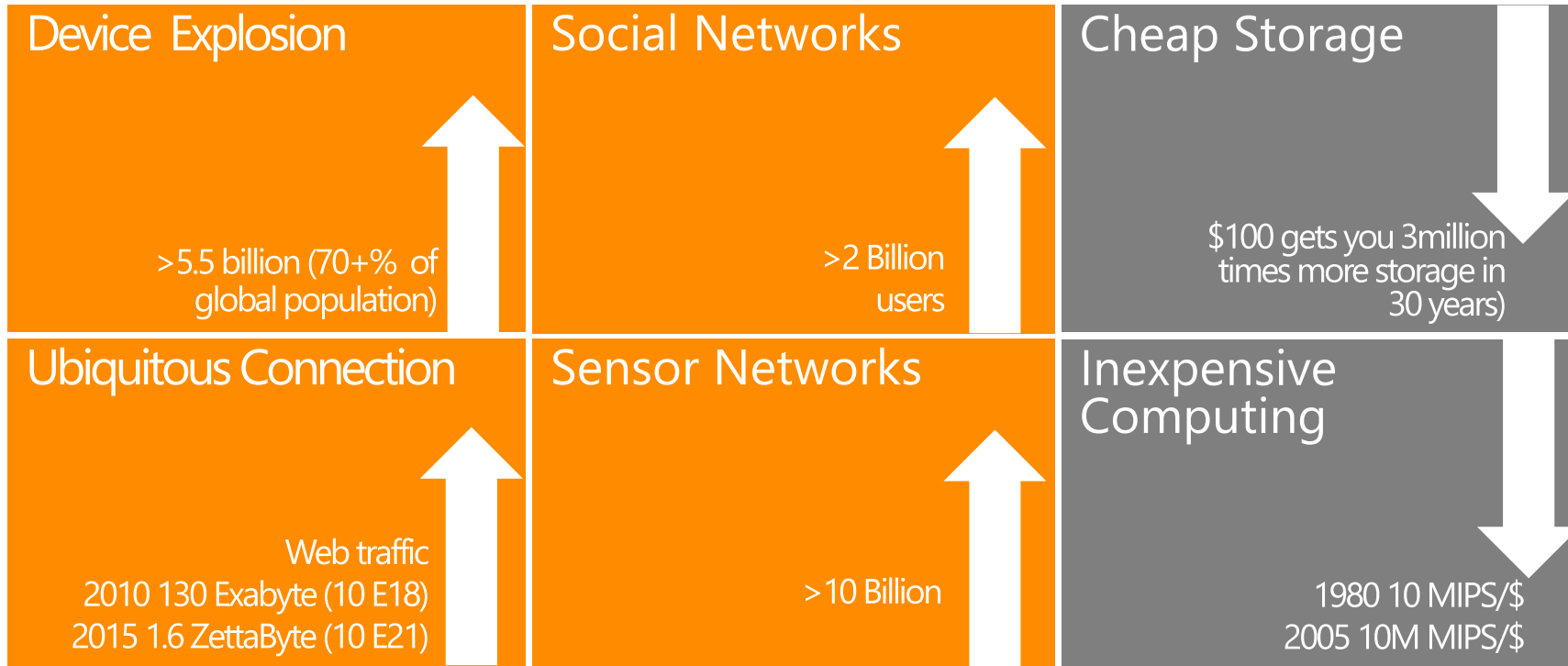
HDInsight



HDInsight



Key Trends



Big Data Analytics

Data
acquisition
& modeling

Collaboration
& visualization

Analysis &
data mining

Dissemination,
sharing,
preservation

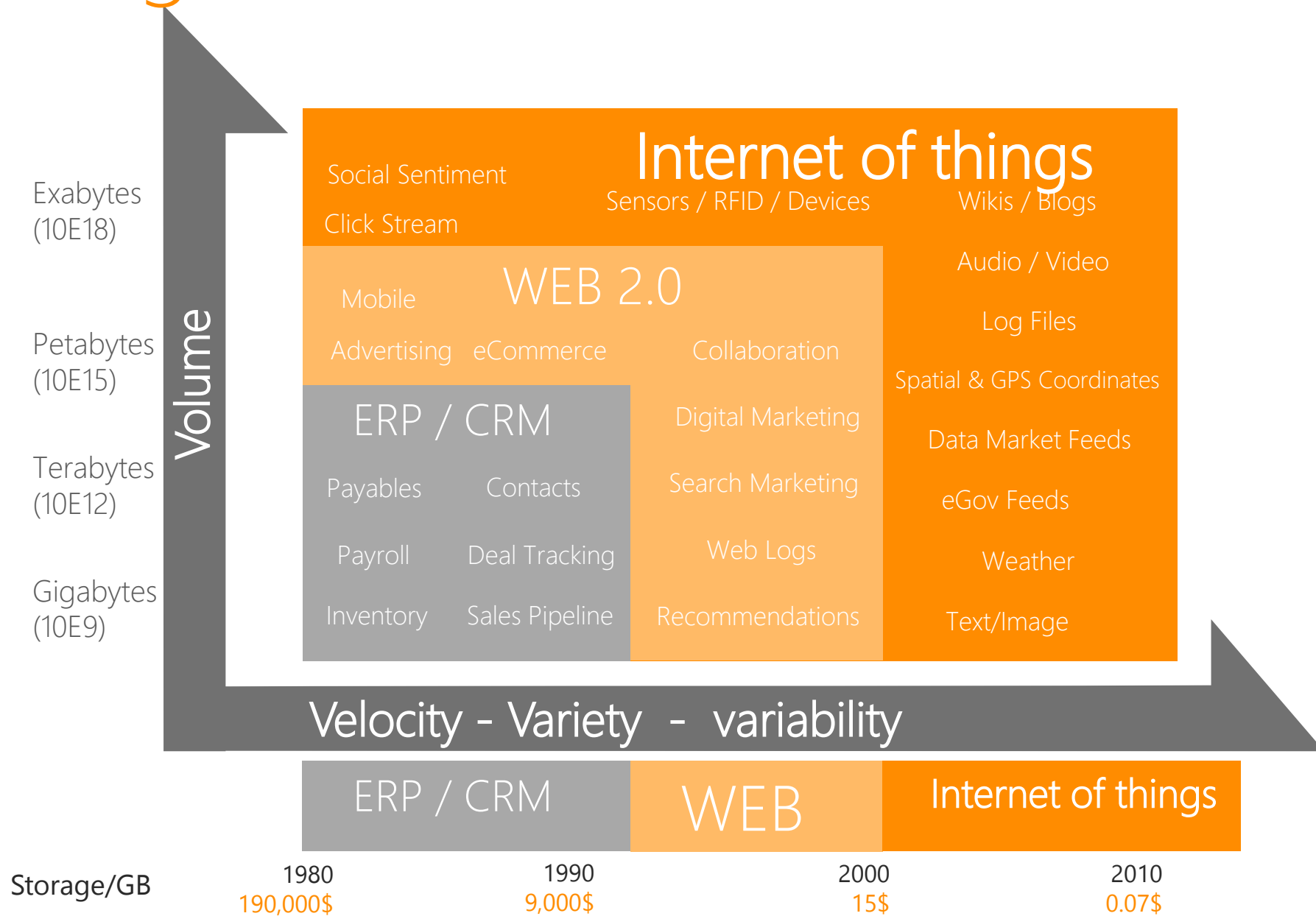
“It takes more time to hand a project from the seismic guys to me to the engineers in production than it does to figure out the oil field plays.”

Geologist,
Major oil and gas company

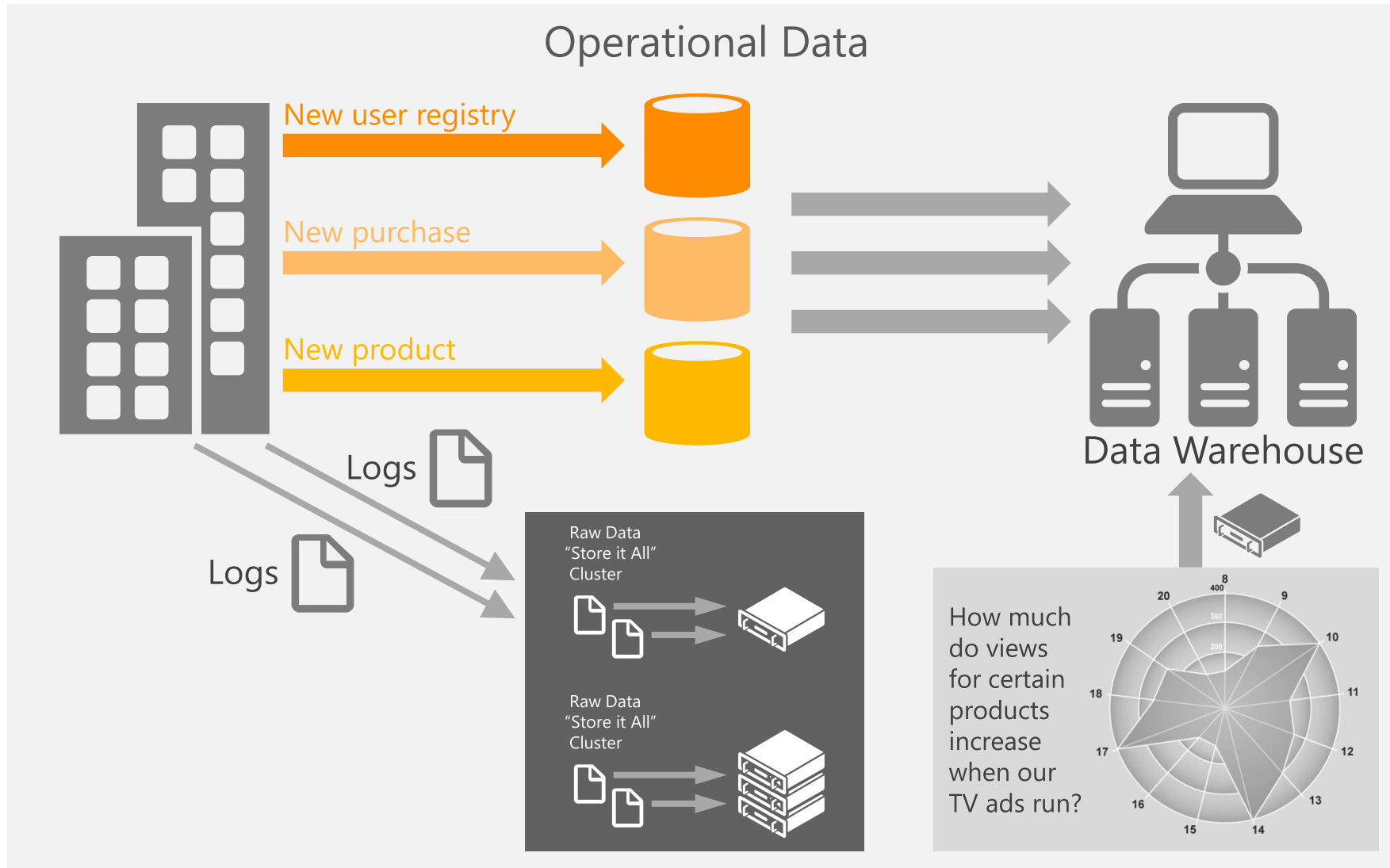
“Our weather model and resulting data sets should be accessible to universities and other institutions.”

Aerospace Development Manager,
U.S. Federal Government

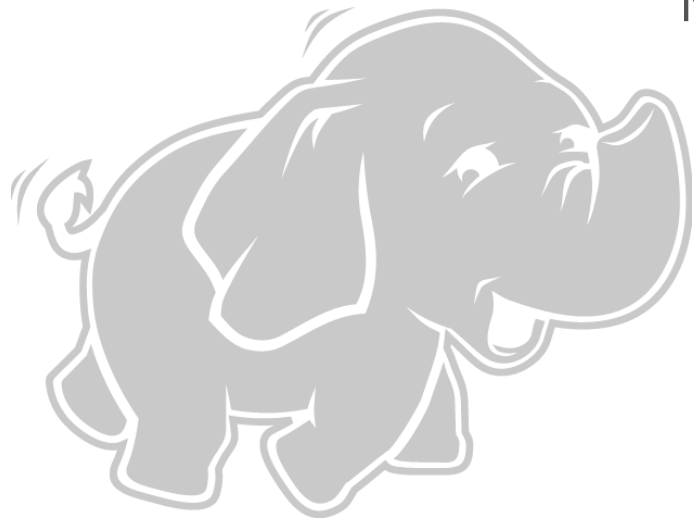
What Is Big Data?



Modern Data Warehousing

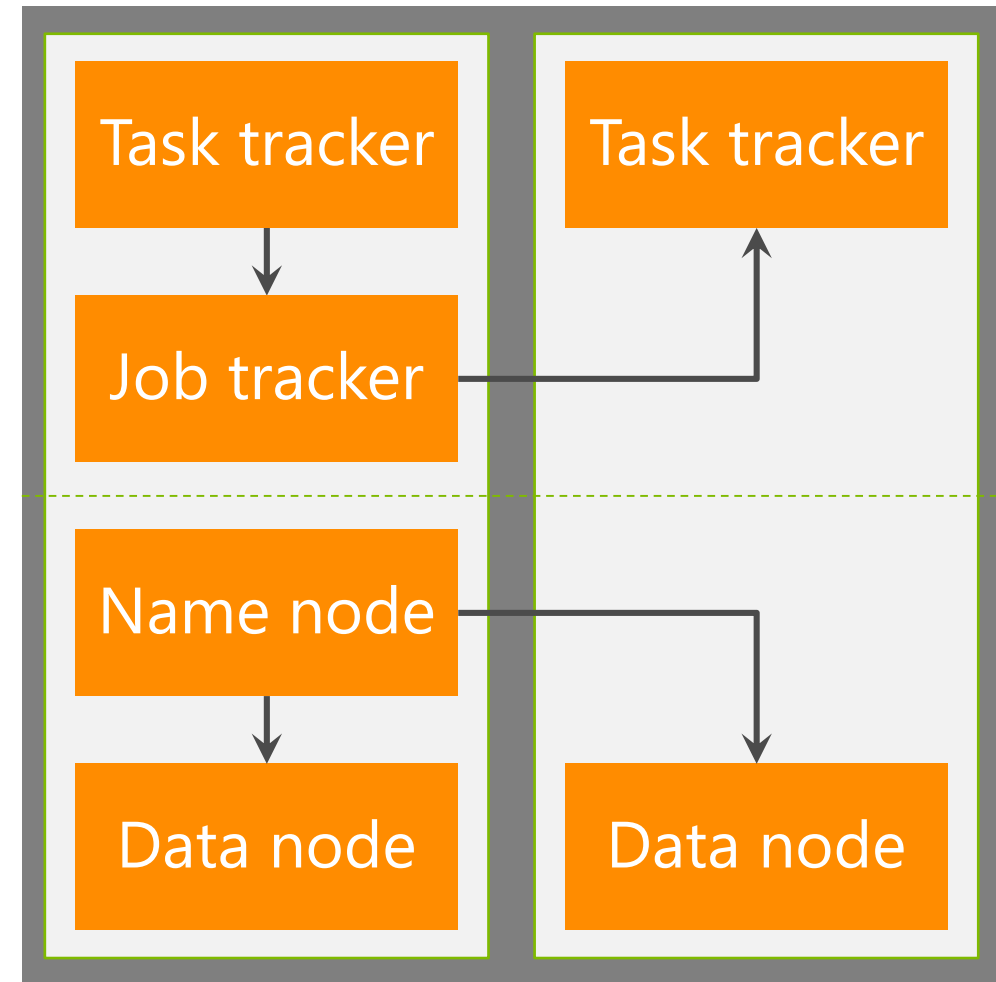


Hadoop Distributed Architecture

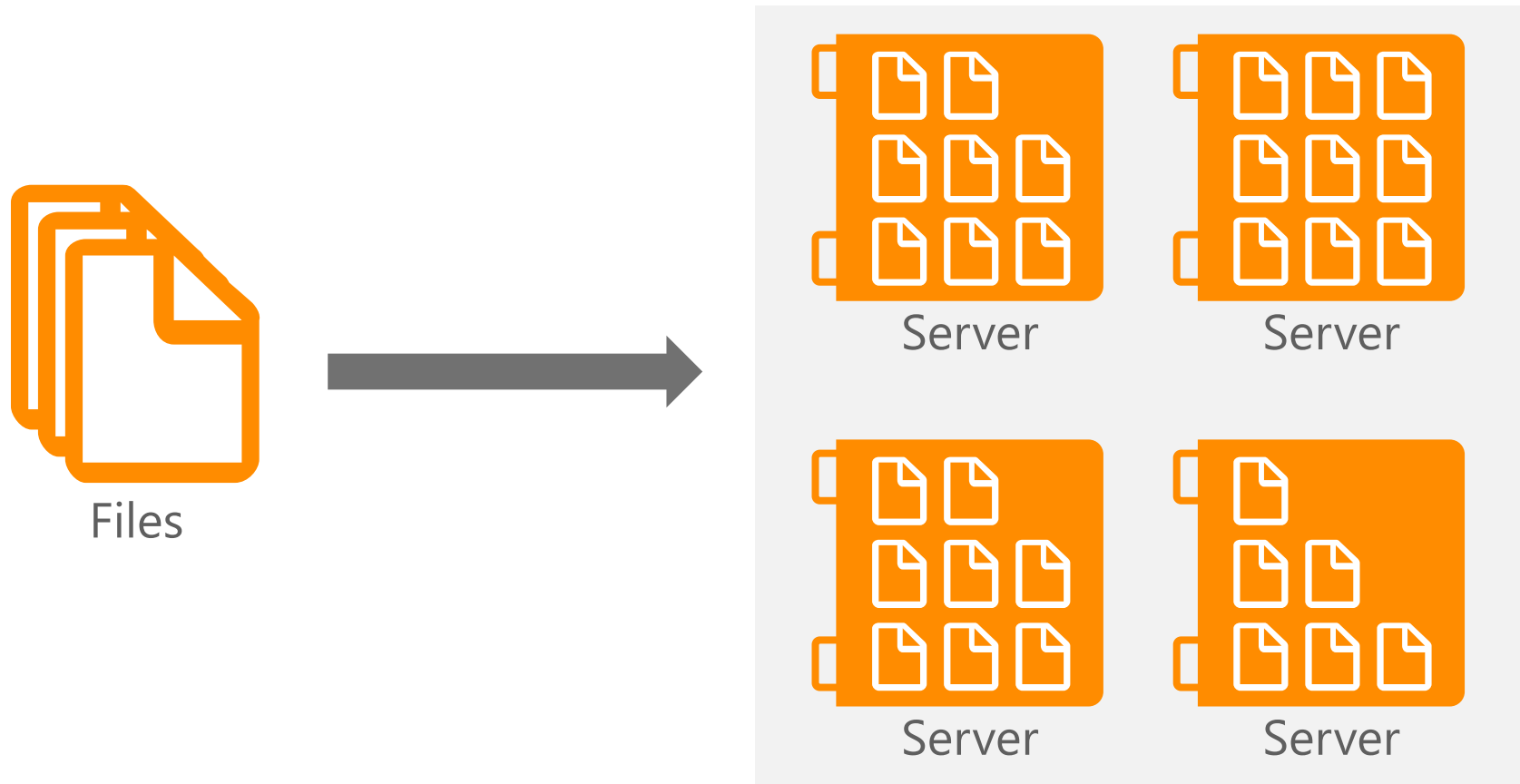


MapReduce
Layer

HDFS
Layer



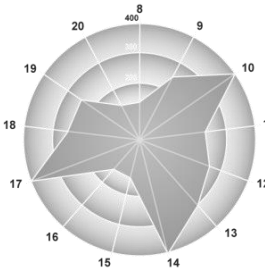
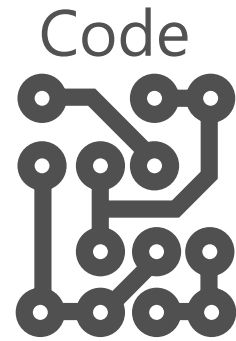
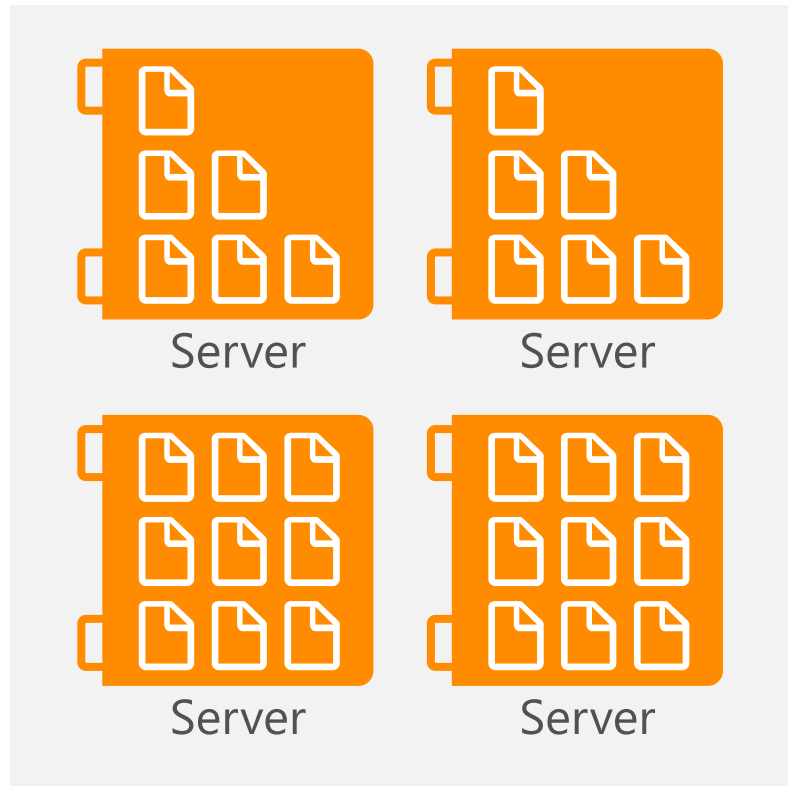
MapReduce: Move Code to the Data



So How Does It Work?

Second, take the processing to the data

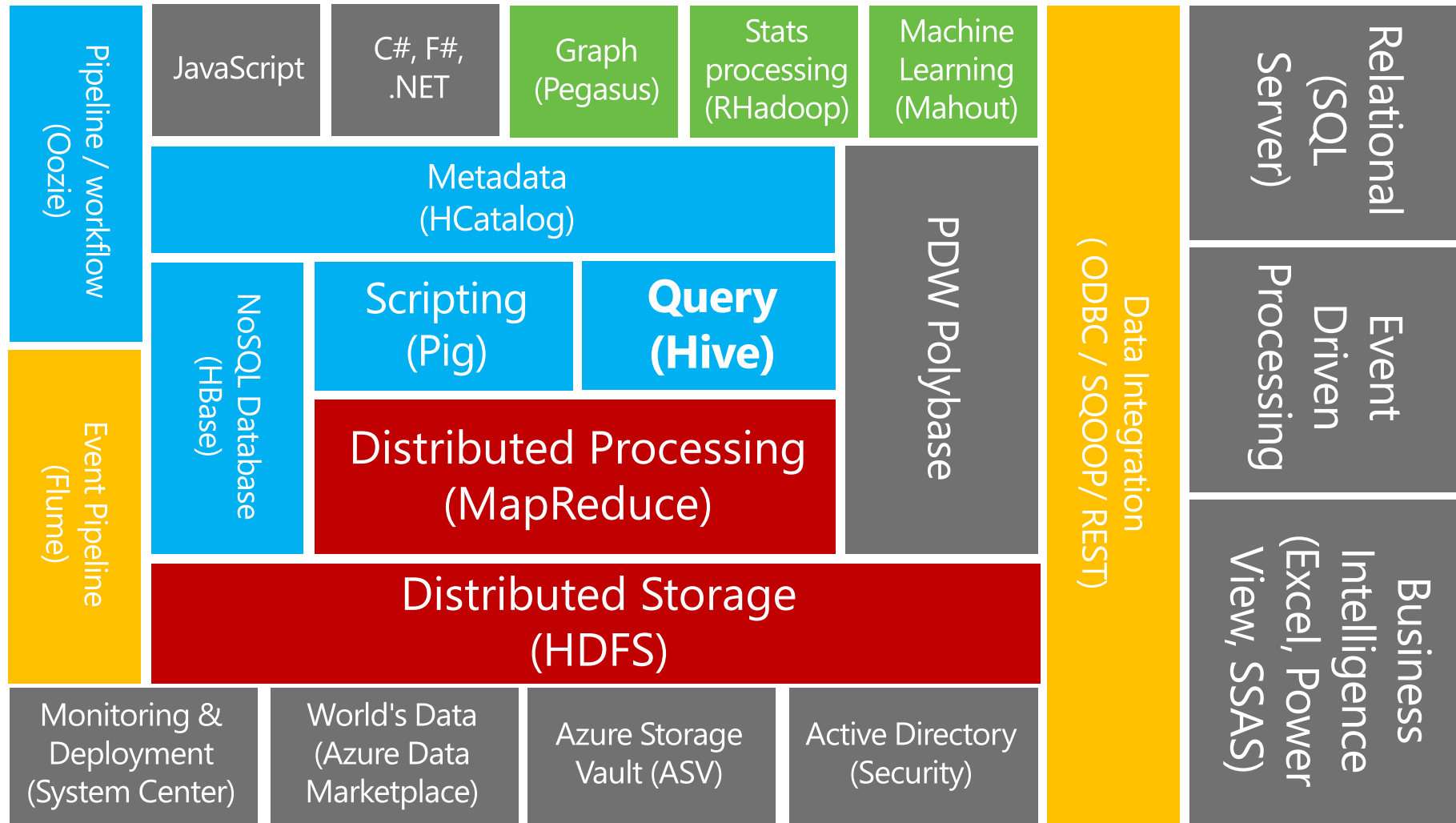
Runtime



```
// 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);
};
```

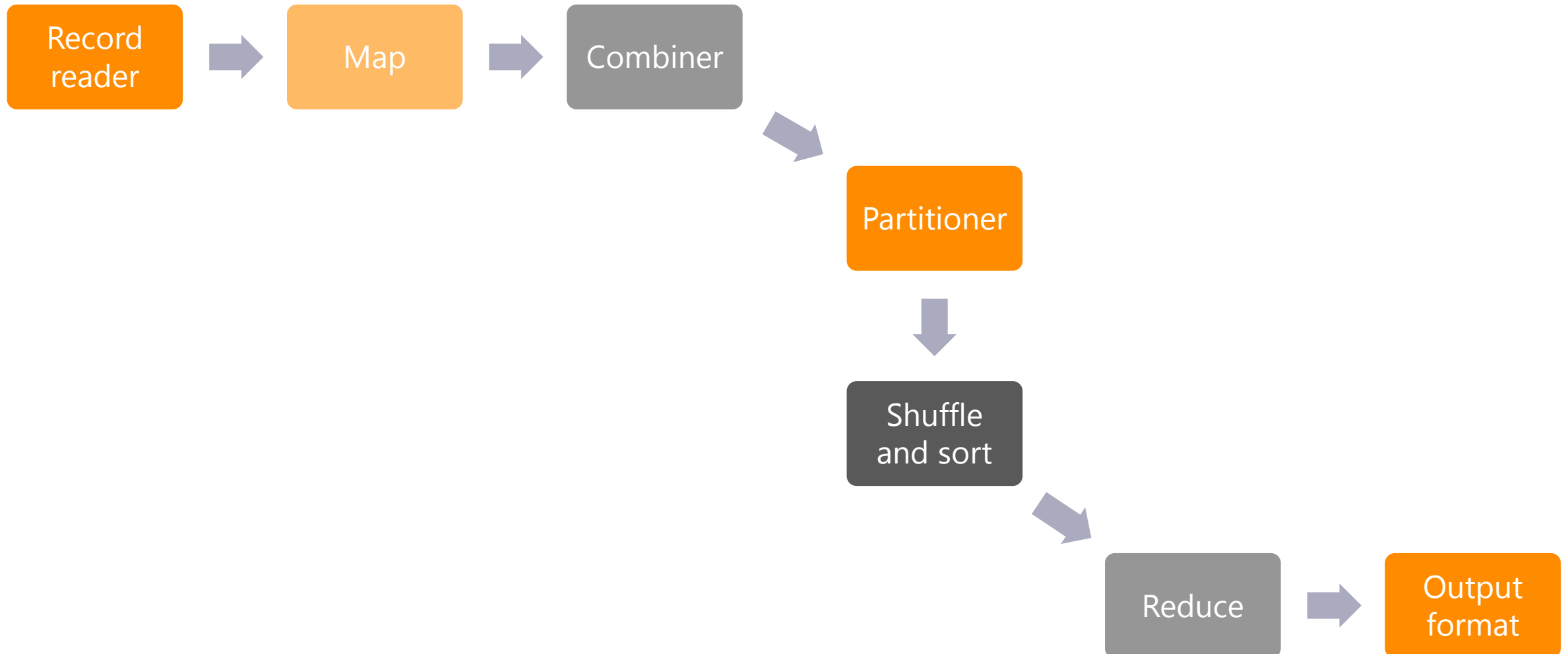
HDInsight and Hadoop Ecosystem



Legend

Red = Core Hadoop
Blue = Data processing
Gray = Microsoft integration points and value adds
Orange = Data Movement
Green = Packages

MapReduce in Hadoop

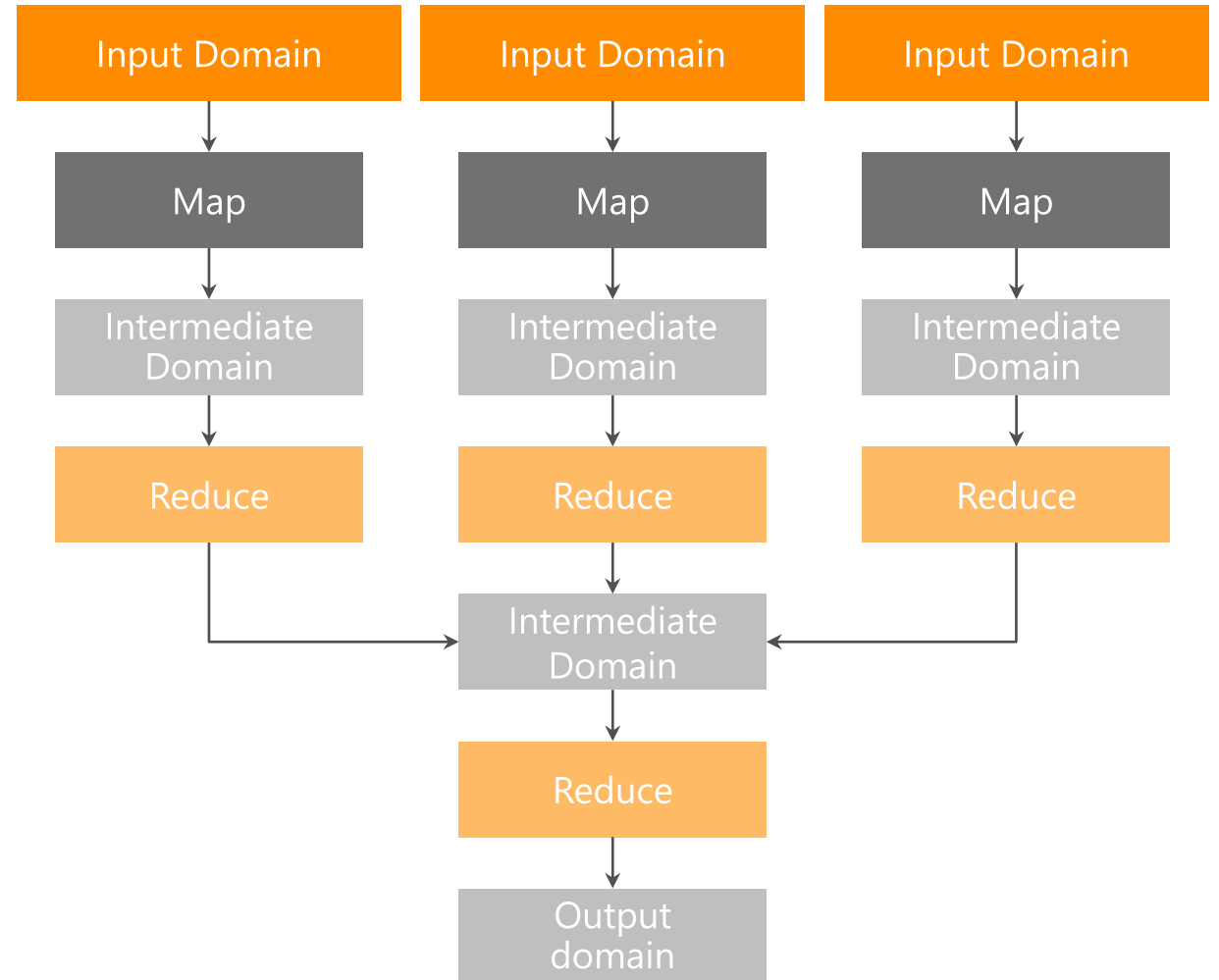


MapReduce: Workflow

A MapReduce *job* usually splits the input data-set into independent chunks which are processed by the *map tasks* in a completely parallel manner

The framework sorts the outputs of the maps, which are then input to the *reduce* tasks

The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks



MapReduce: Driver

```
public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    String[] otherArgs =
        new GenericOptionsParser(conf, args).getRemainingArgs();
    if (otherArgs.length != 2) {
        System.err.println("Usage: CommentWordCount <in> <out>");
        System.exit(2);
    }

    Job job = new Job(conf, "StackOverflow Comment Word Count");
    job.setJarByClass(CommentWordCount.class);
    job.setMapperClass(WordCountMapper.class);
    job.setCombinerClass(IntSumReducer.class);
    job.setReducerClass(IntSumReducer.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
    FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
}
```

MapReduce: Mapper

```
public void map(Object key, Text value, Context context)
    throws IOException, InterruptedException {

    // Parse the input string into a nice map
    Map<String, String> parsed = MRDPUtils.transformXmlToMap(value.toString());

    // Grab the "Text" field, since that is what we are counting over
    String txt = parsed.get("Text");

    // .get will return null if the key is not there
    if (txt == null) {
        // skip this record
        return;
    }

    // Unescape the HTML because the data is escaped.
    txt = StringEscapeUtils.unescapeHtml(txt.toLowerCase());

    // Remove some annoying punctuation
    txt = txt.replaceAll("'", ""); // remove single quotes (e.g., can't)
    txt = txt.replaceAll("[^a-zA-Z]", " "); // replace the rest with a space

    // Tokenize the string by splitting it up on whitespace into
    // something we can iterate over,
    // then send the tokens away
    StringTokenizer itr = new StringTokenizer(txt);
    while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
        context.write(word, one);
    }
}
```

MapReduce: Reducer

```
public void reduce(Text key, Iterable<IntWritable> values,  
    Context context) throws IOException, InterruptedException {  
    int sum = 0;  
    for (IntWritable val : values) {  
        sum += val.get();  
    }  
  
    result.set(sum);  
    context.write(key, result);  
}
```

MapReduce Summary

- Analogous to GROUP BY in SQL
- Usually combiners can be an effective way to minimize network IO
- Based on the dataset opportunity to use a custom partitioner to balance the reducers



Programming HDInsight

Existing
ecosystem

Hive, Pig, Mahout, Cascading, Scalding, Scoobi, Pegasus...

Microsoft .NET

C#, F# Map/Reduce, LINQ to Hive, Microsoft .NET
management clients

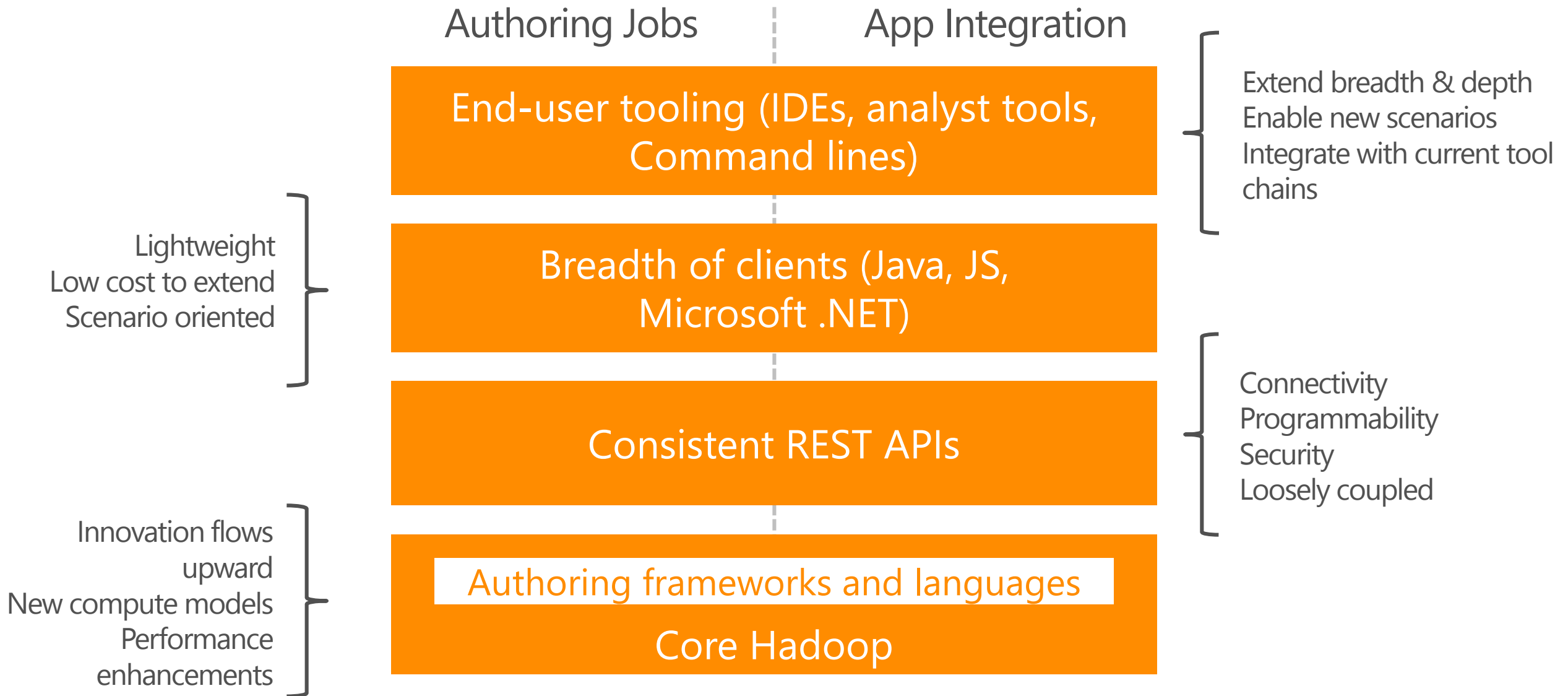
JavaScript

JavaScript Map/Reduce, browser hosted console, Node.js
management clients



DevOps / IT
Pros

PowerShell, cross-platform CLI tools

Building Developer Experiences



RDBMS vs. Hadoop

	RDBMS	HADOOP
Data size	Gigabytes (<i>Terabytes</i>)	Petabytes (<i>Exabytes</i>)
Access	Interactive and Batch	Batch
Updates	Read / Write many times	Write once, Read many times
Structure	Static Schema	Dynamic Schema
Integrity	High (ACID)	Low
Scaling	Nonlinear	Linear
		

Reference: Tom White's Hadoop: The Definitive Guide

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



Polybase

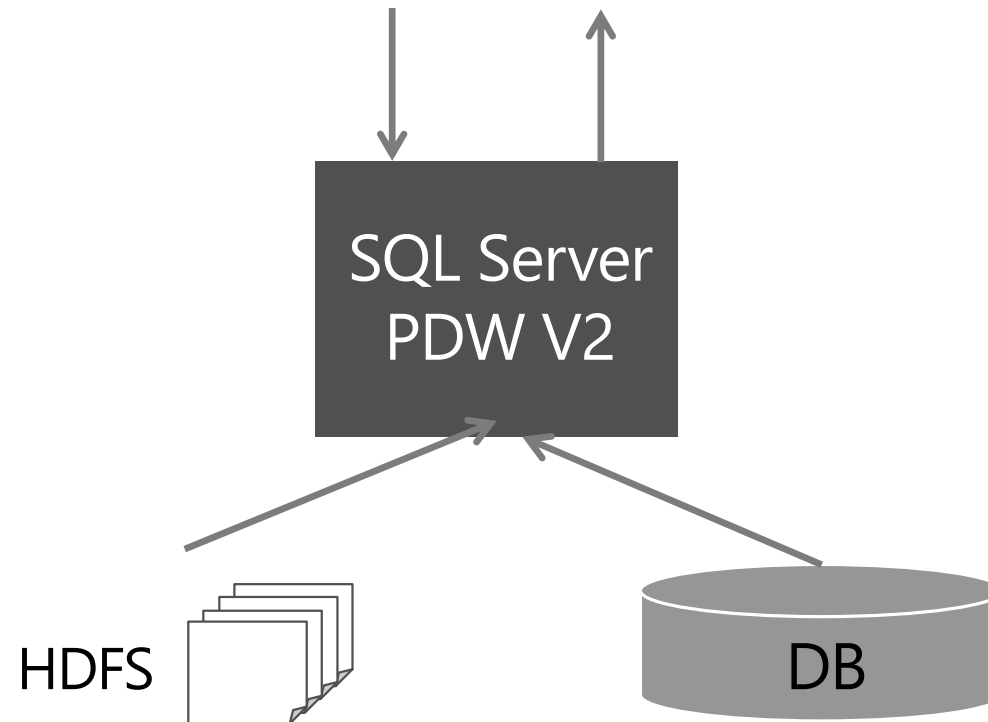


Polybase

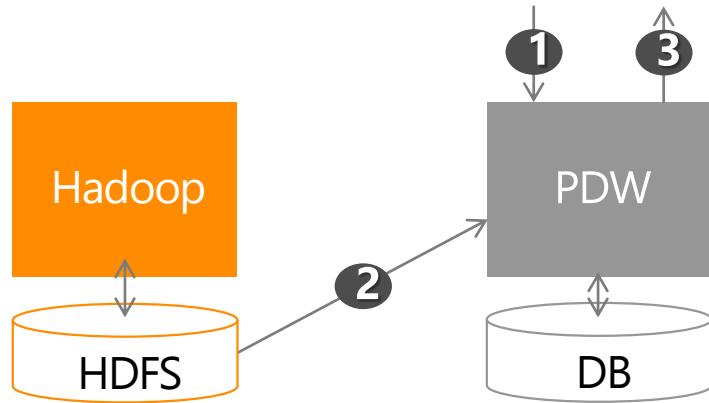


Polybase

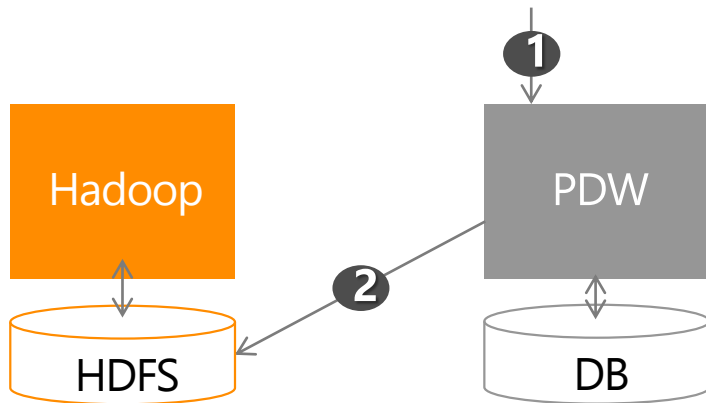
SQL Server PDW querying HDFS data, in-situ



Polybase in PDW V2



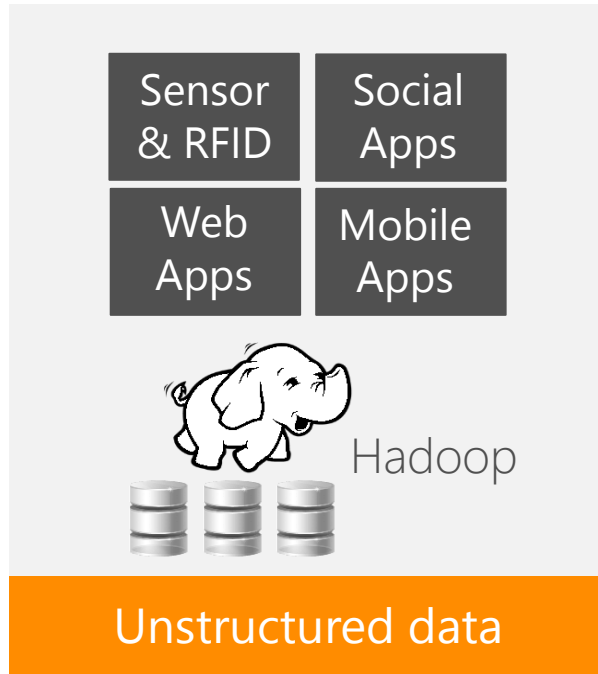
(a) PDW query in, results out



(b) PDW query in, results stored in HDFS

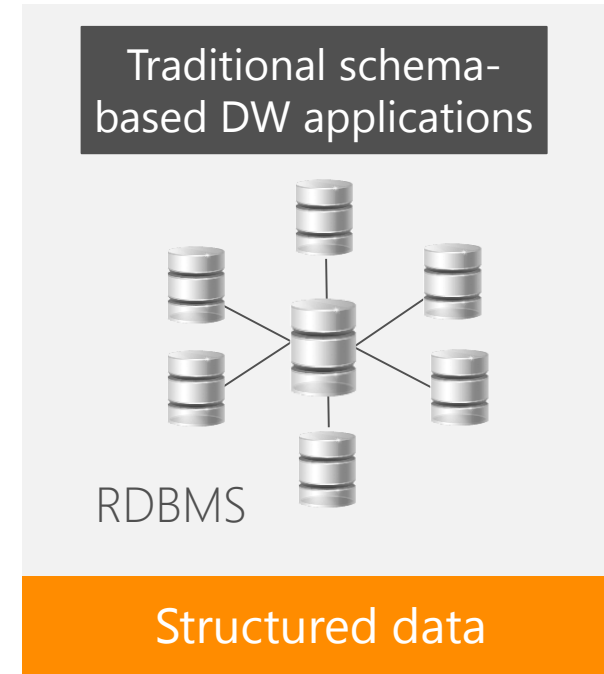
1. HDFS as new distribution type for PDW tables
2. Parallel loads directly from HDFS into PDW
3. Fully parallel, bidirectional CTAS between PDW and HDFS
4. All query processing done by PDW nodes
5. Both Linux and Windows clusters supported

Native Query Across Hadoop and PDW



Increasingly massive amounts of unstructured data driven by new sources

How to overcome the "impedance mismatch"



At the same time, vast amounts of corporate data and data sources, and the bulk of their data analysis

Polybase addresses this challenge for advanced data analytics by allowing native query across PDW and Hadoop, integrating structured and unstructured data

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

Native Query Across Hadoop and PDW

Creating "External Tables"

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

1

Indicates
"External" Table

2

Required location of
Hadoop cluster and file

3

Optional Format Options associated
with data import from HDFS

- Internal representation of data residing in Hadoop/HDFS (delimited text files only)
- High-level permissions required for creating external tables
 - **ADMINISTER BULK OPERATIONS & ALTER SCHEMA**
- Different than 'regular SQL tables': essentially read only (no DML support)

Native Query Across Hadoop and PDW

Querying Unstructured Data

1. Querying data in HDFS and displaying results in table form (using external tables)
2. Joining data from HDFS with relational PDW data

Example – Creating external table 'ClickStream':

```
CREATE EXTERNAL TABLE ClickStream(url varchar(50), event_date date, user_IP  
varchar(50)), WITH (LOCATION = 'hdfs://MyHadoop:5000/tpch1GB/employee.tbl',  
FORMAT_OPTIONS (FIELD_TERMINATOR = '|' ));
```

Query Examples

Text file in HDFS with | as field delimiter

- 1

```
SELECT top 10 (url) FROM ClickStream where user_IP = '192.168.0.1'
```

 Filter query against data in HDFS
- 2

```
SELECT url.description FROM ClickStream cs, Url_Description url  
WHERE cs.url = url.name and cs.url='www.cars.com';
```

 Join data coming from files in HDFS
(Url_Description is a second text file in HDFS)
- 3

```
SELECT user_name FROM ClickStream cs, Users u WHERE  
cs.user_IP = u.user_IP and cs.url='www.microsoft.com';
```

 Join data from HDFS with relational PDW table
(Users is a distributed PDW table)

Native Query Across Hadoop and PDW

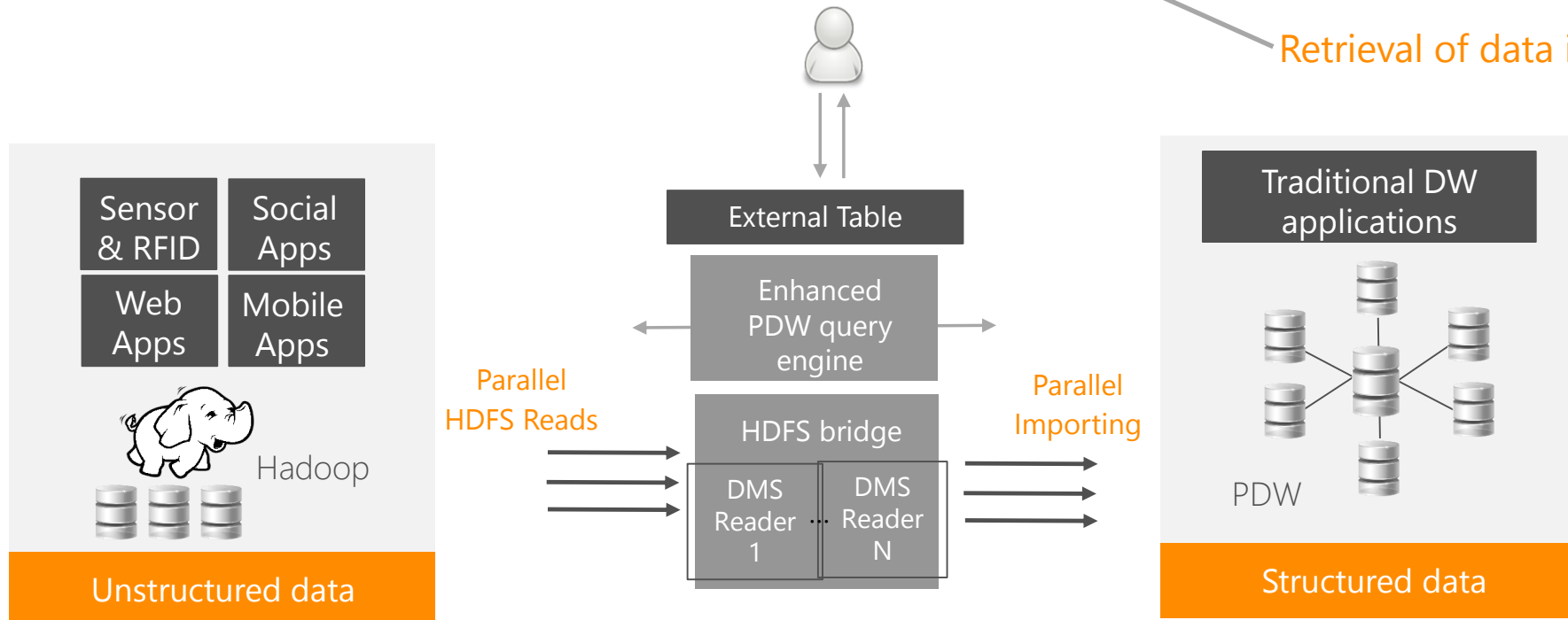
Parallel Data Import from HDFS into PDW

Persistently storing data from HDFS in PDW tables

Fully parallelized via CREATE TABLE AS SELECT (CTAS) with external tables as source table and PDW tables (either distributed or replicated) as destination

```
CREATE TABLE ClickStream_PDW WITH DISTRIBUTION = HASH(url)
AS SELECT url, event_date, user_IP FROM ClickStream
```

Retrieval of data in HDFS "on-the-fly"

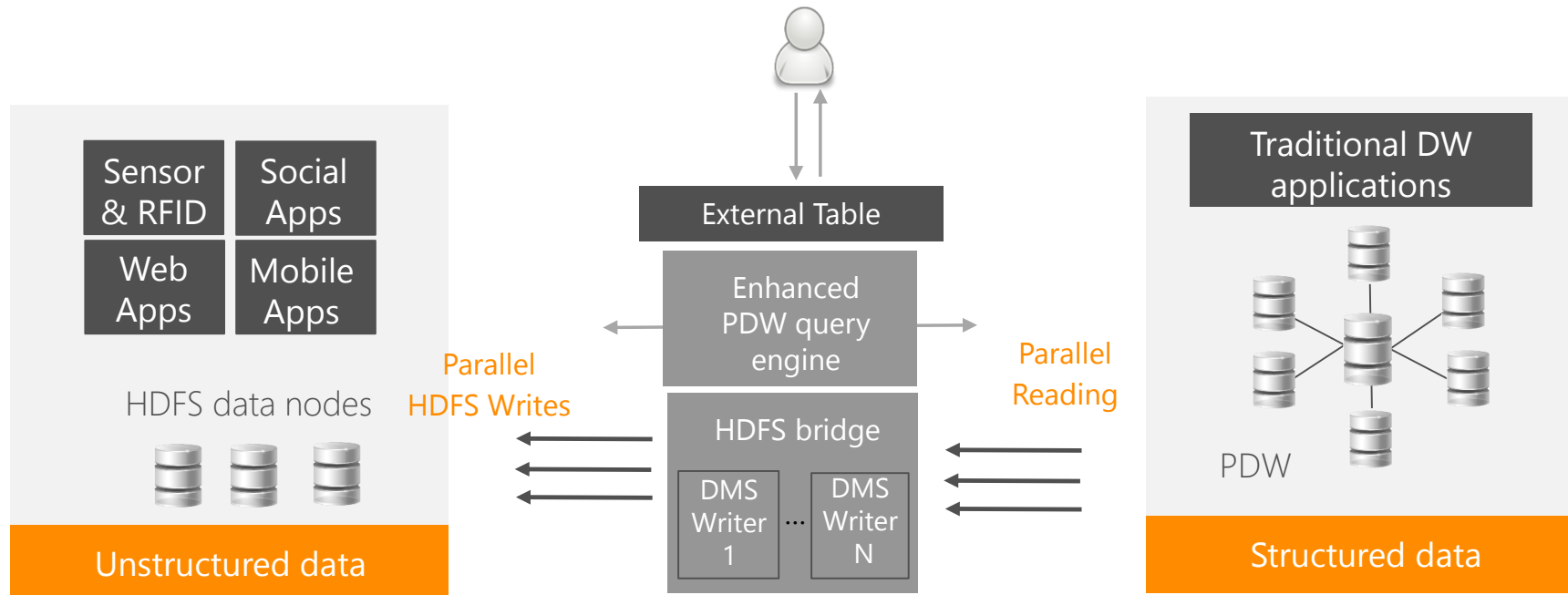


Native Query Across Hadoop and PDW

Parallel Data Export from PDW into HDFS

- Fully parallelized via CREATE EXTERNAL TABLE AS SELECT (CETAS) with external tables as destination table and PDW tables as source
- 'Round-trip of data' possible with first importing data from HDFS, joining it with relational data, and then exporting results back to HDFS

```
CREATE EXTERNAL TABLE ClickStream (url, event_date, user_IP)  
WITH (LOCATION = 'hdfs://MyHadoop:5000/users/outputDir', FORMAT_OPTIONS  
(FIELD_TERMINATOR = '|')) AS SELECT url, event_date, user_IP FROM ClickStream_PDW
```



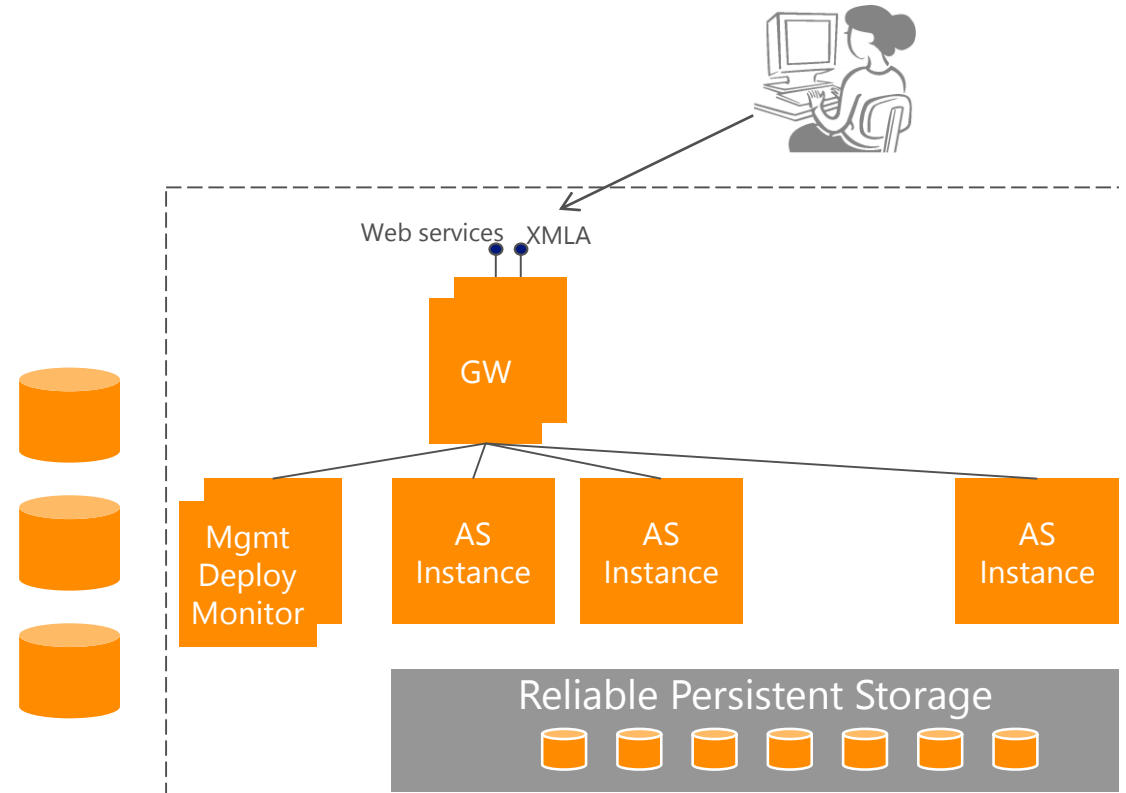
In-Memory for big data analytics

Interactive Analytics over "Big Data"



Clusters

- SQL Server Analysis Services scaled out to very large data volumes
- Sourced from "Big Data" sources, e.g.
 - Hadoop, Isotope, etc.
 - Enterprise data sources (SQL Server, Oracle, SAP, etc.)
- Built upon the In-Memory Analytics engine
 - In-memory, column-store, 10x compression
- Deployment vehicles: Box, Appliance, Cloud
- Customers:
 - Skype, Klout, Halo 4, UBS, AdCenter, Windows Update





StreamInsight

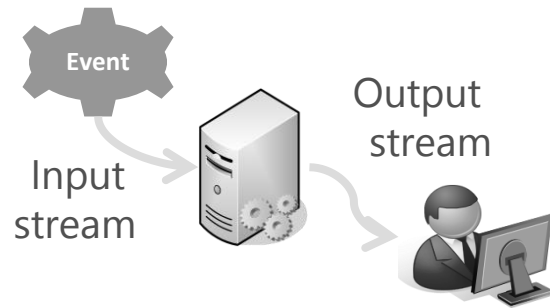


StreamInsight



StreamInsight

Managing Streaming Data In-Memory



edgenet
Sell Easier | Sell More

- StreamInsight is a Complex Event Processing tool
- Ideal for continuous streaming data e.g. web click streams, stock trading data or smart grid data
- Runs completely *in-memory*

Customer benefits

- Low latency with sub-zero processing of large event stream
- Rich developer experience with .NET 4.0 and Visual Studio
- Ease of management
- Flexible deployment – from embedded devices, regional hubs or centralized location

“With our current solution, we’re about one hour delayed from a real-time response to incoming events. But with SQL Server 2008 R2 StreamInsight, we expect to cut that time to seconds..”

Christopher Barwick,
Enterprise Analytics Team Lead

Analyzing Streaming Data in the Cloud

StreamInsight on Windows Azure



Clusters



StreamInsight on Azure is a cloud scale service for complex event-processing

Ideal for analyzing streaming data either born on the cloud, or globally distributed

Customer benefits

- Insights from data in motion
- Elastic scale out in the cloud
- Simplified management via built-in connectivity
- Low TCO of cloud service

Complete and Consistent Data Platform



SQL Server 2014

Mission-critical performance
Faster insights from any data
Platform for hybrid cloud

Development

Management

Data

Identity

Virtualization

Call to action



Download SQL Server 2014 CTP2

Stay tuned for availability

www.microsoft.com/sqlserver

