Microsoft

Polybase

Agenda

- What is Polybase
- Agnostic Architecture
- Parallel Data Transfer
- Hybrid Queries
- Split Query Execution
- Configuration

Polybase: What’s all the fuss about?


ing Flour
SUGAR

Baking a big data cake

Ingredients

- Data
- Questions
- Desire for answers

Question 3:
How do I
query it?

Answer:
You tell me!

No one store to rule them all…

- There is only data, questions and answers
- Business users do not care for
  - Technology
  - Complexity
- Business users do care about
  - Costs (especially opex)
  - Getting answers (quickly)
  - Staying competitive

Users want to query all data across types and locations

Reduce Costs

Drivers

- Retain existing skills
- Reduce complexity
- Use commodity kit

Solutions

- Use familiar tools
- KISS principle
- Avoid proprietary systems

Getting Answers

Goals

- Simple integration
- High performance
- Low latency
- Query across <u>all</u> data

Solutions

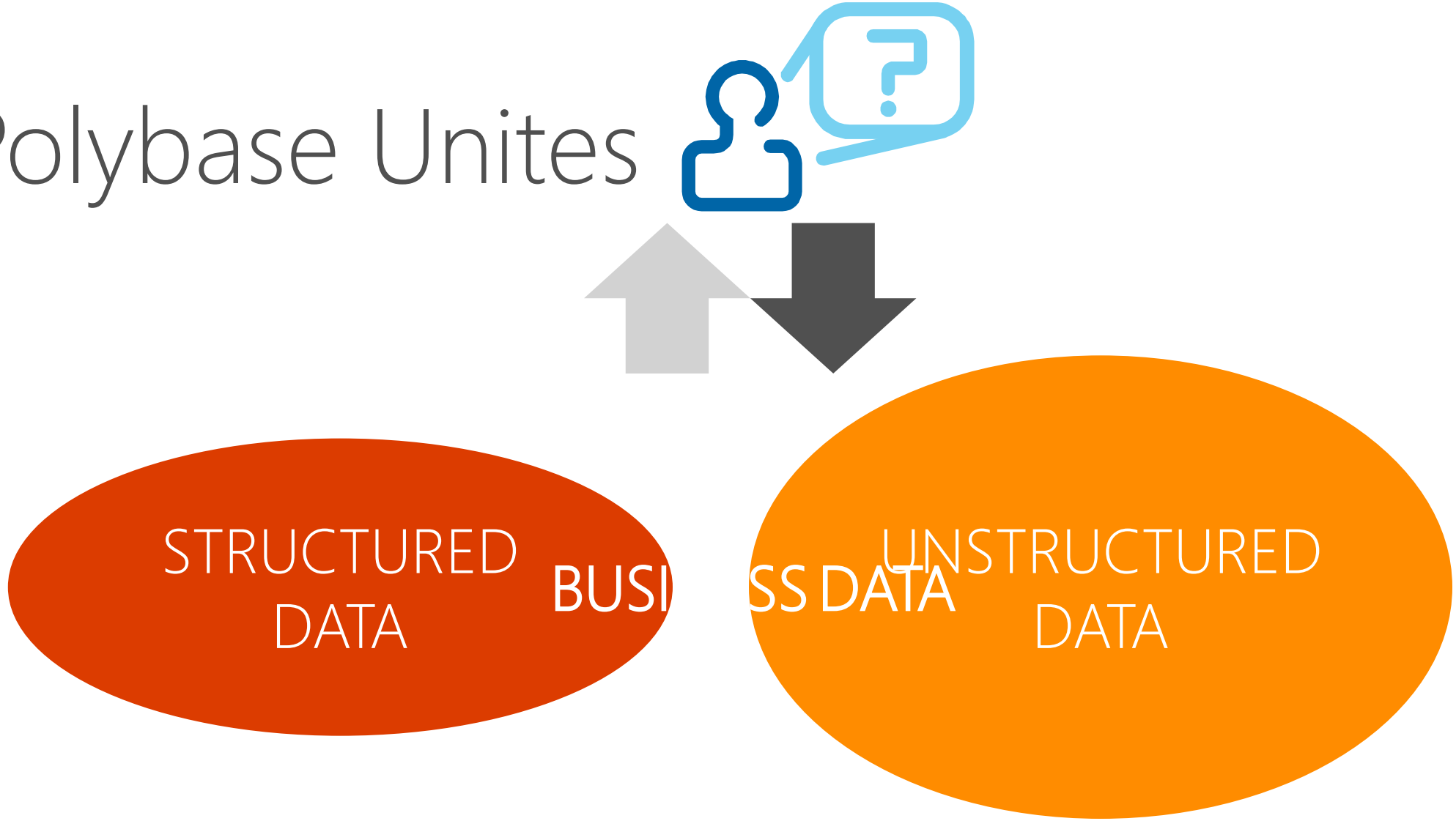
- Dynamic solution
- Scalable with demand
- Minimize data movement
- Distributed engine

Staying Competitive

- Ask bigger questions
- Faster time to insight
- Complete the picture
- Lower cost of curiosity
- Perform new analysis
- Avoid vendor lock-in

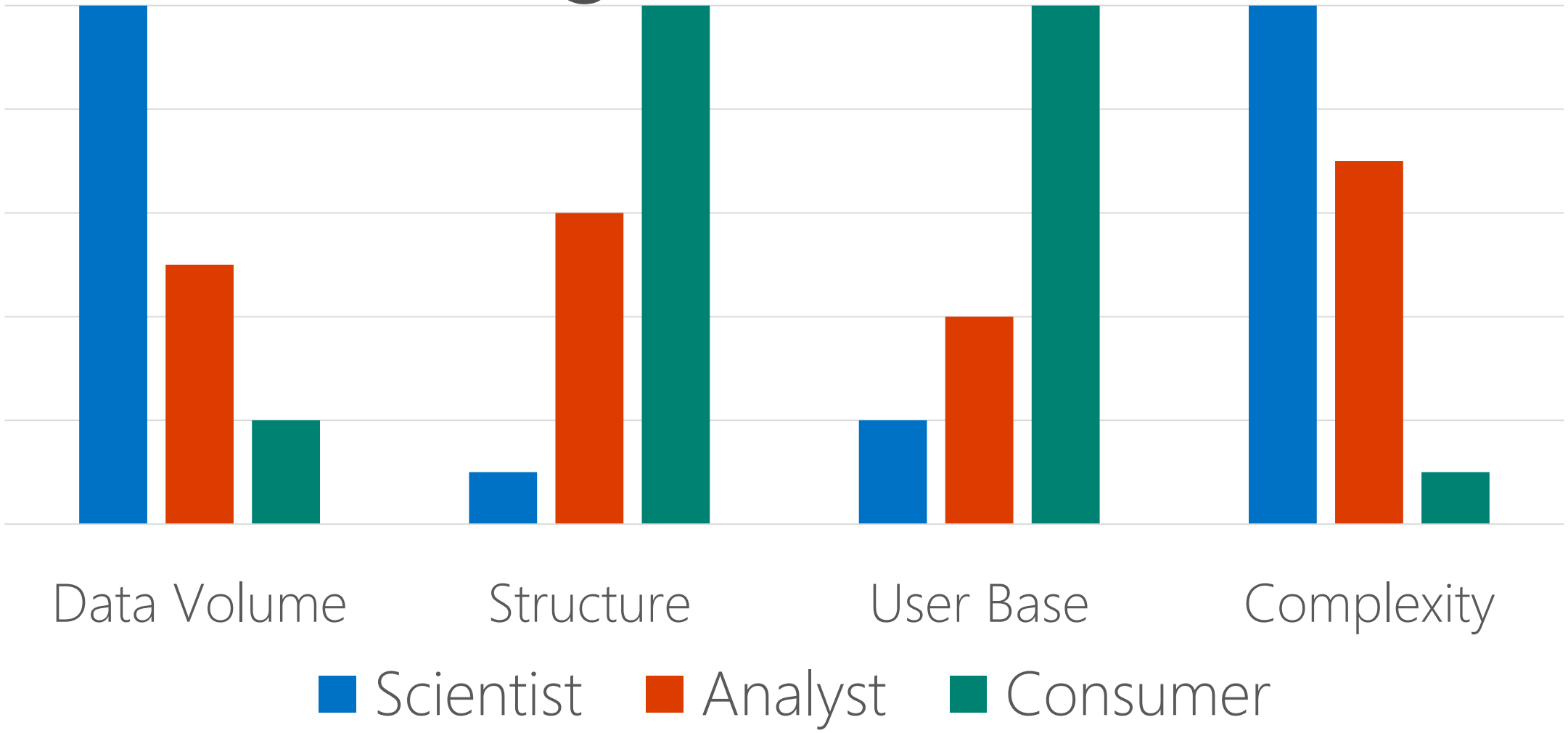
- Hadoop integration
- Iterative development
- Query across all datasets
- Flexible platform
- Use new functionality
- Agnostic architecture
  - Format
  - Structure
  - Location

Polybase Unites



...for a better together world of analytics

Characterising User Personas



What's the sweet spot for Polybase?

	Consumer	Analyst	Scientist
Data Volume	Medium to Low	Reasonable	High -> Huge
Degree of Structure	Very High	Some	Low ->None
Number of Users	Very High	Medium	Low
Transformation Complexity	Low	Medium to High	High
Analytics Complexity	Low	Medium	Very High

Partial fit for Polybase today
Structure possibly absent on data
Good option for data delivery & transform

Polybase Builds The Bridge

- Just-in-Time data integration
  - Across relational and non-relational data
  - High performance parallel architecture
  - Fast, simple data loading
- Best of both worlds
  - Uses computational power at source for both relational data & Hadoop
  - Opportunity for new types of analysis
- Uses existing analytical skills
  - Familiar SQL semantics & behaviour
- Query with familiar tools
  - SSDT

Polybase = run time integration

Includes Power BI

Agnostic Architecture

Polybase is agnostic
=
No vendor lock in

Polybase supports Hadoop on Linux & Windows

Polybase integrates with the cloud

Polybase supports HDInsight in APS & external Hadoop clusters

Loosely Coupled Architecture

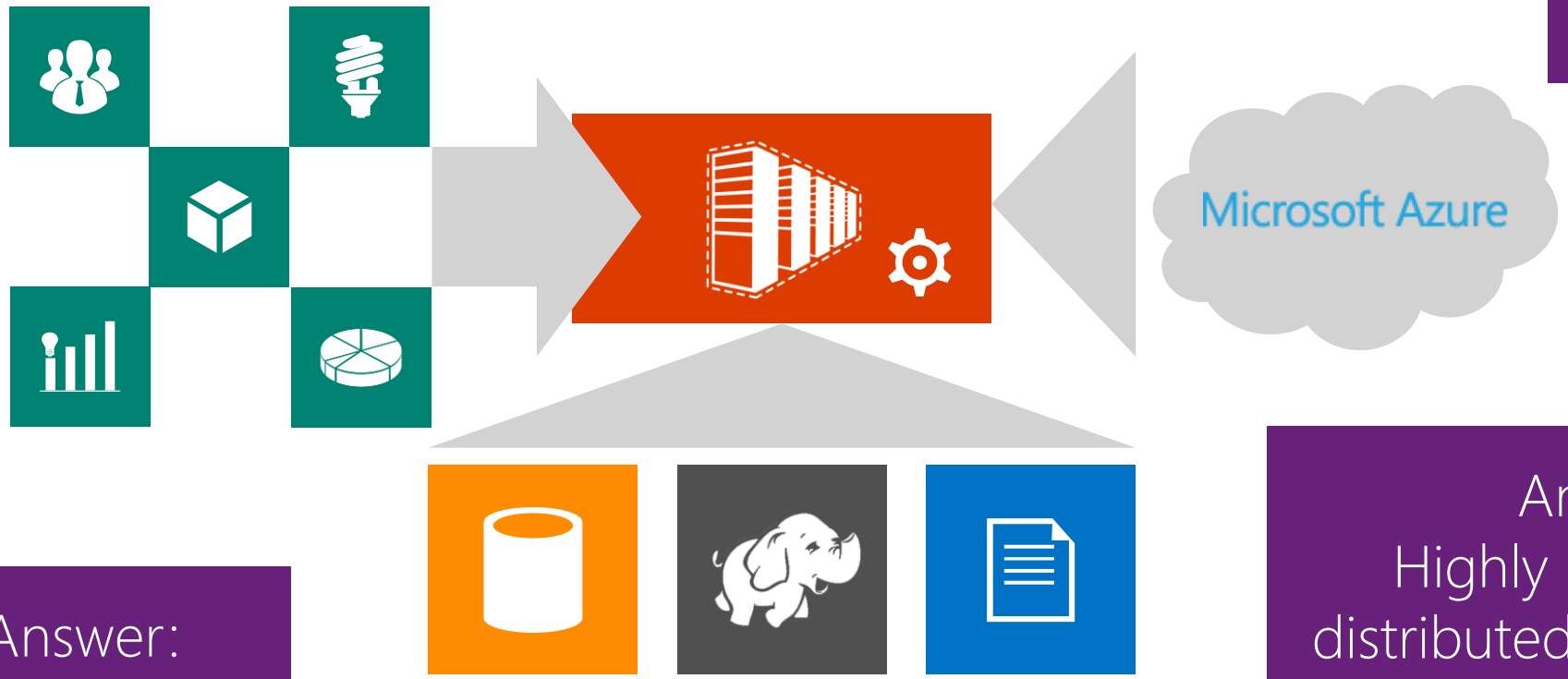
Late Binding Consequences

- Data may change between executions
- Data may change during execution
- Errors identified at run time

All "By Design"
Helps Polybase
keep its agnostic
architecture

So what is Polybase?

Answer:
Component of the PDW Region in APS



Answer:
Unique Innovative Technology

Answer:
Seamless Integration

Answer:
Highly parallelised distributed query engine accessing heterogeneous data via SQL

What are the goals of Polybase?

The goals are…

- Make data accessible
- Make it fast
- Make it easy
- Make data access format agnostic
- Parallel Data Transfer
- Runtime integration using SQL

Any Data in Any format

Deployment Choices

Hortonworks Hadoop On Windows (External)

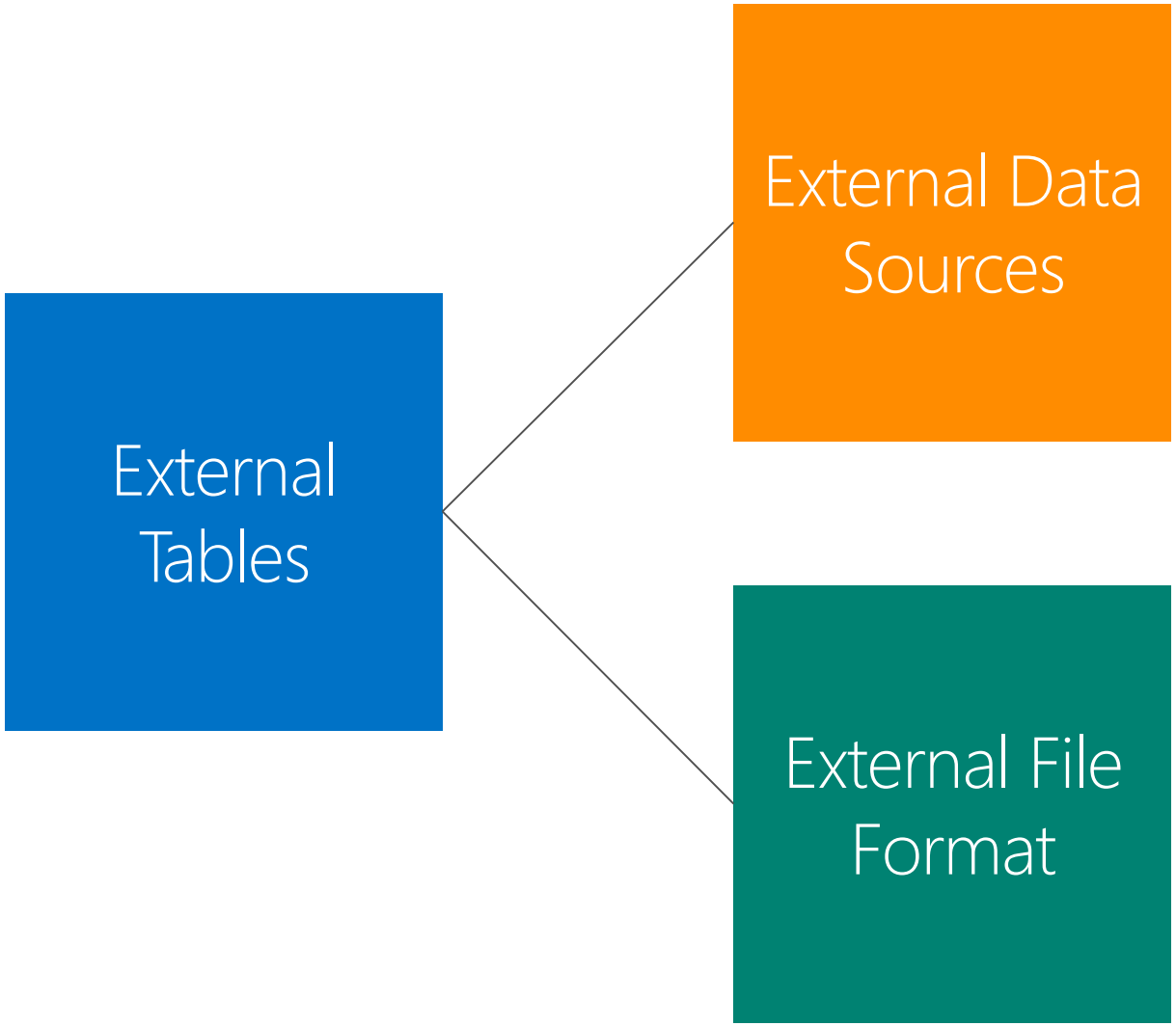
Hortonworks Hadoop On Linux (External)

Cloudera CDH On Linux (External)

HDInsight On APS (Internal)

HDInsight On WASB (External)

Staying Agnostic



External Tables

- Metadata used to describe external data
- Enables data access outside the PDW region
- Never hold data
- Do not delete data when dropped

Behaviour of an external table is in PDW is very similar to Hive external tables

External Tables – Catalog Views

Logical table in shell database (control node)

- sys.external_tables
- sys.tables

Create External Table

```
CREATE EXTERNAL TABLE [dbo].[Sales]
([ProductKey]       int   NOT NULL
,[StoreKey]         int   NOT NULL
,[DateKey]          int   NOT NULL
,[CustomerKey]      int   NOT NULL
,[PromotionKey]     int   NOT NULL
,[OrderQuantity]    int   NOT NULL
,[UnitPrice]        money NOT NULL
,[SalesAmount]      money NOT NULL
)
```

Syntax for External Tables has been enriched for AU1

External Tables in AU0.5 (Deprecated)

```
WITH
(LOCATION='hdfs://ip_address:port/files/Sales'
,FORMAT_OPTIONS (FIELD_TERMINATOR    = '|'
                ,STRING_DELIMITER    = ''
                ,DATE_FORMAT         = ''
                ,REJECT_TYPE         = VALUE
                ,REJECT_VALUE        = 0
                ,USE_TYPE_DEFAULT    = False
                )
);
```

AU 0.5
location & format of data were tightly bound to external table

External Tables in AU1.0

```
WITH
(LOCATION='hdfs://filepath_or_directory'
,DATA_SOURCE          = MyDataSourceName
,FILE_FORMAT          = MyFileFormatName
,REJECT_TYPE          = VALUE
,REJECT_VALUE         = 0
,REJECT_SAMPLE_VALUE = 1000
);
```

AU 1
location & format
have been
de-coupled

External Table Creation

STEP ID	OPERATION	LOCATION	DISTRIBUTION	ROW COUNT	START TIM
0	OnOperation	Control	Unspecified	-1	4/2/2014 1:

```
CREATE EXTERNAL TABLE [TPCH].[dbo].[HDI_Warehouse] ([w_warehouse_sk] INT NOT NULL, [w_warehouse_id] CHAR(16) COLLATE dat
[w_warehouse_sq_ft] INT, [w_street_number] CHAR(10) COLLATE database_default, [w_street_name] VARCHAR(60) COLLATE databa
CHAR(10) COLLATE database_default, [w_city] VARCHAR(60) COLLATE database_default, [w_county] VARCHAR(30) COLLATE databas
database_default, [w_country] VARCHAR(20) COLLATE database_default, [w_gmt_offset] DECIMAL(5, 2)) WITH (DATA_SOURCE = [H
[HDI_HadoopRegion_TextPipe],REJECT_TYPE = PERCENTAGE,REJECT_VALUE = 1,REJECT_SAMPLE_VALUE = 1 )
```

1	OnOperation	Control	Unspecified	-1	4/2/2014 1:

```
EXEC [TPCH].[sys].[sp_addextendedproperty] @name=N'pdw_physical_name', @value=N'_85a3605fafa248a78aa2632e6235989d', @lev
@level1name=N'HDI_Warehouse'
```

2	OnOperation	Control	Unspecified	-1	4/2/2014 1:

```
EXEC [TPCH].[sys].[sp_addextendedproperty] @name=N'pdw_distribution_type', @value=N'External', @level0type=N'SCHEMA', @l
```

3	ExternalStatisticsOperation	Control	Unspecified	-1	4/2/2014 1:

```
<Empty>
```

4	OnOperation	Control	Unspecified	-1	4/2/2014 1:

```
UPDATE STATISTICS [TPCH].[dbo].[HDI_Warehouse] WITH ROWCOUNT = [ROWCOUNT_TEMP_ID_51], PAGECOUNT = [PAGECOUNT_TEMP_ID_51]
```

External table extended properties

Extended Properties	Definition
pdw_physical_name	Internal mapping name of the external table exposed via sys.pdw_table_mappings
pdw_distribution_type	Determines table geometry. In this case identifies the table as an external table. Value is therefore=N'External'. Other values are Distributed and Replicated

External Table Limitations

- No Insert / Update / Delete functionality
  - Select
  - Bulk Import and Export
- No integration with external metadata sources
  - HCatalog

Same for Hive

Duplication of metadata

External Data Source

Puts the "Poly" into Polybase

- Introduces the concept of a location type
- Opens the door for integrating other sources
- Allows other optional configurations to be set
- sys.external_data_sources is the catalog view

Dropping an external data source impacts all external tables that depend on it invalidating them

External tables must be dropped and re-created once invalidated

External Data Source – Hadoop Cluster

```
CREATE EXTERNAL DATA SOURCE MyHadoopDataSource
WITH
(TYPE          = HADOOP
,LOCATION      = 'hdfs://NameNode_URI[:port]'
,JOB_TRACKER_LOCATION = 'JobTracker_URI[:port]'
)
;
```

Setting the Job Tracker Location enables the generation of MapReduce jobs against the Hadoop Cluster

Hadoop Region & External Data Source

- To ensure data flows over the IB network additional host names have been created
- It is *imperative* that you use the correct naming for NameNode and Job Tracker

Hadoop Region & External Data Source

```
CREATE EXTERNAL DATA SOURCE HDI_HadoopRegion_DataSource
WITH (   TYPE          = HADOOP
      ,  LOCATION = 'hdfs://HTUKIA-C-HHN01:8020'
      ,  JOB_TRACKER_LOCATION = 'HTUKIA-C-HHN01:50300'
      );
```

Use these names

External Data Source – WASB[S]

```
CREATE EXTERNAL DATA SOURCE MyAzureDataSource
WITH
(TYPE          = HADOOP
,LOCATION      =
'wasb[s]://[container@]account_name.blob.core.windows.net/path'
)
;
```

Type is still Hadoop

No Job Tracker Parameter available
No Pushdown Predicate for WASB[S] Source

External File Format Enhancements

- RCFiles
- Data Compression
- De-coupled from External table
- sys.external_file_formats

External File Format - RCFiles

- RCFile format now supported
- RC = Record Columnar
- Key/value pairs
- Used for storing data in columnar format
- SERDE methods access RCFiles
- Other serde methods can be installed

SERDE
=
Serialization
DeSerialization

External File Format - Limitations

- Row Terminator is fixed as \n
- Encoding is also fixed : UTF8
- Compression choice may be limited by format

External File Format – Delimited Text

```
CREATE EXTERNAL FILE FORMAT MyTextFileFormat
WITH
(FORMAT_TYPE          =    DELIMITEDTEXT
,FORMAT_OPTIONS       =    (FIELD_TERMINATOR= '|'
                           ,STRING_DELIMITER= ','
                           ,DATE_FORMAT= ymd
                           ,USE_TYPE_DEFAULT= TRUE
                           )
,DATA_COMPRESSION  =
'org.apache.hadoop.io.compress.DefaultCodec'
| 'org.apache.hadoop.io.compress.GzipCodec'
);
```

External File Format – Hadoop RC File

```
CREATE EXTERNAL FILE FORMAT MyRCFileFormat
WITH
(FORMAT_TYPE        = RCFILE
,SERDE_METHOD       =
'org.apache.hadoop.hive.serde2.columnar.LazyBinaryColumnarSerDe'
| 'org.apache.hadoop.hive.serde2.columnar.ColumnarSerDe'
,DATA_COMPRESSION =
'org.apache.hadoop.io.compress.DefaultCodec'
);
```

External File Format Notes

LazyBinaryColumnarSerDe is significantly faster and more efficient than ColumnarSerDe

Data Compression not designed for the Hadoop Region as the IB connectivity is so fast

Data Compression more beneficial for external clusters using low speed networks

Parallel Data Transfer

Parallel Transfer Concepts

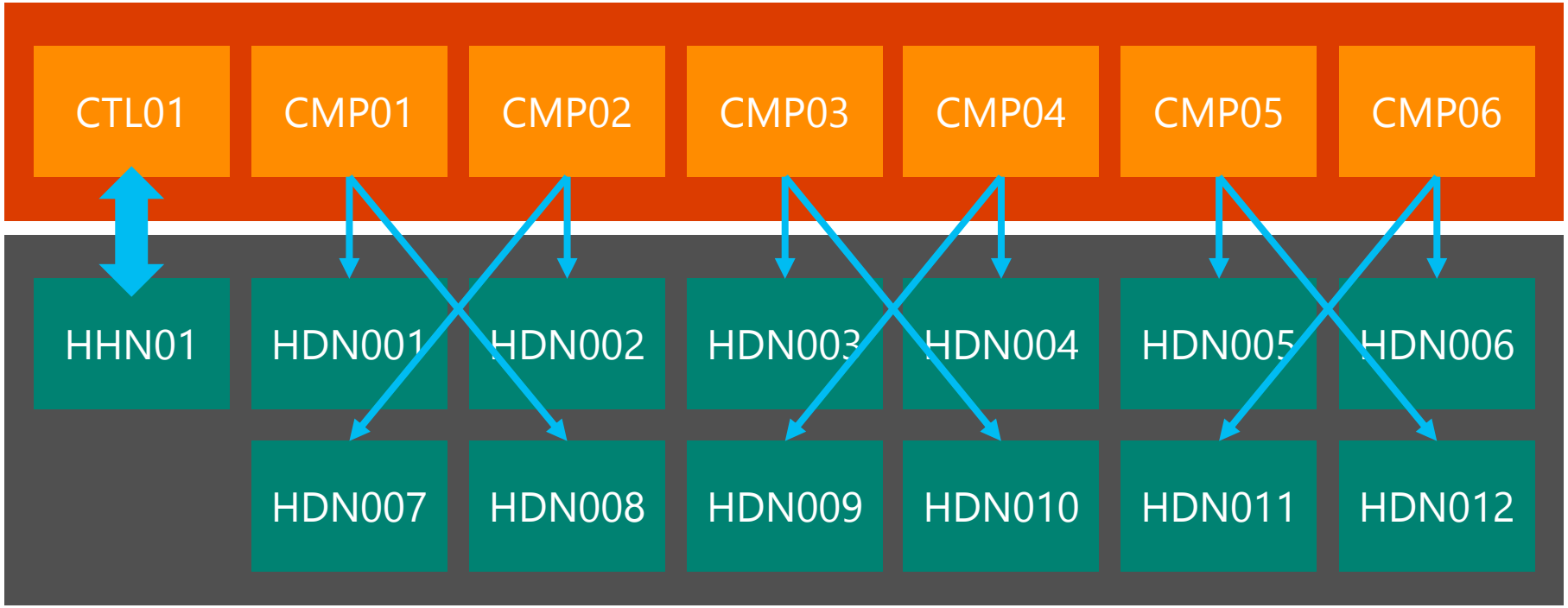
Maximise Throughput

- Every compute node in PDW sees every data node in Hadoop
- Ensure direct connections are established between all scale out nodes of PDW & Hadoop

Balanced Execution

- Ensure all nodes are equally busy when reading and writing data

Maximising Throughput



Polybase & DMS

Implemented as a DMS extension

A new bridge component has been added to DMS

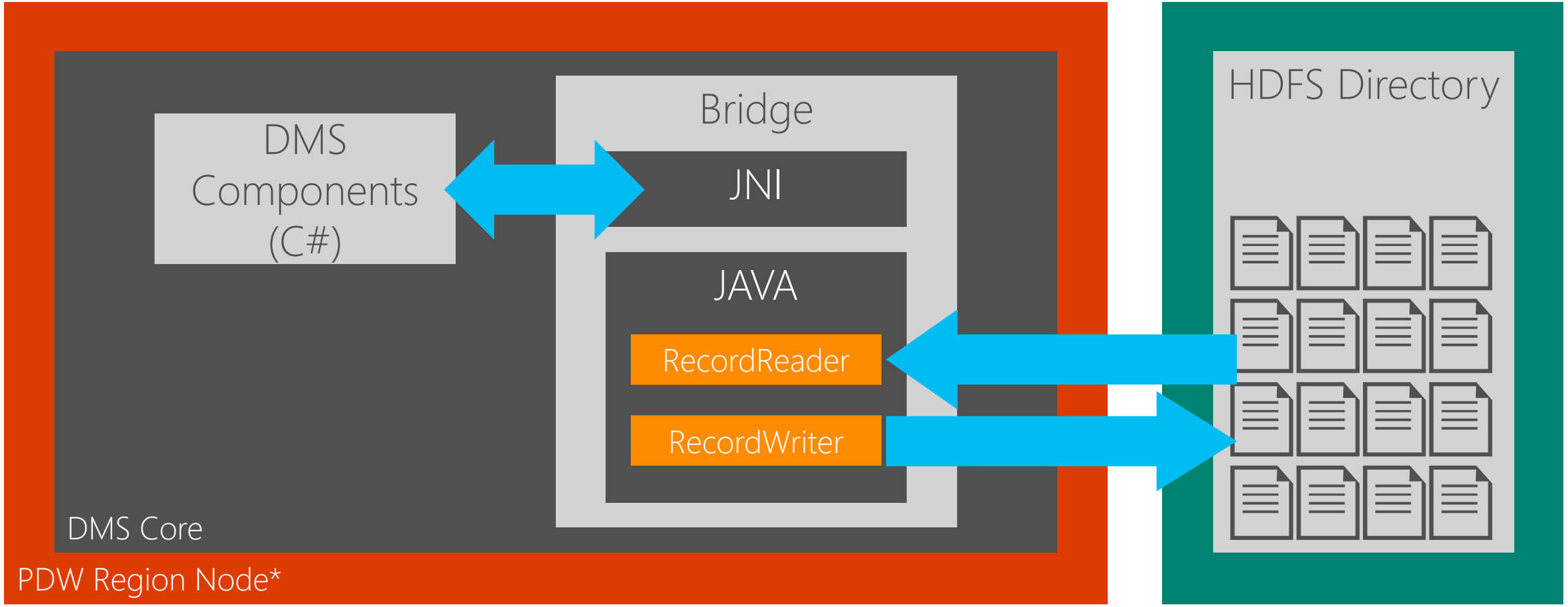
Bridge supports pluggable interfaces for heterogeneous data access

Bridge abstracts the complexity of Hadoop

A Java Native Interface (JNI) layer provides interoperability with the rest of DMS

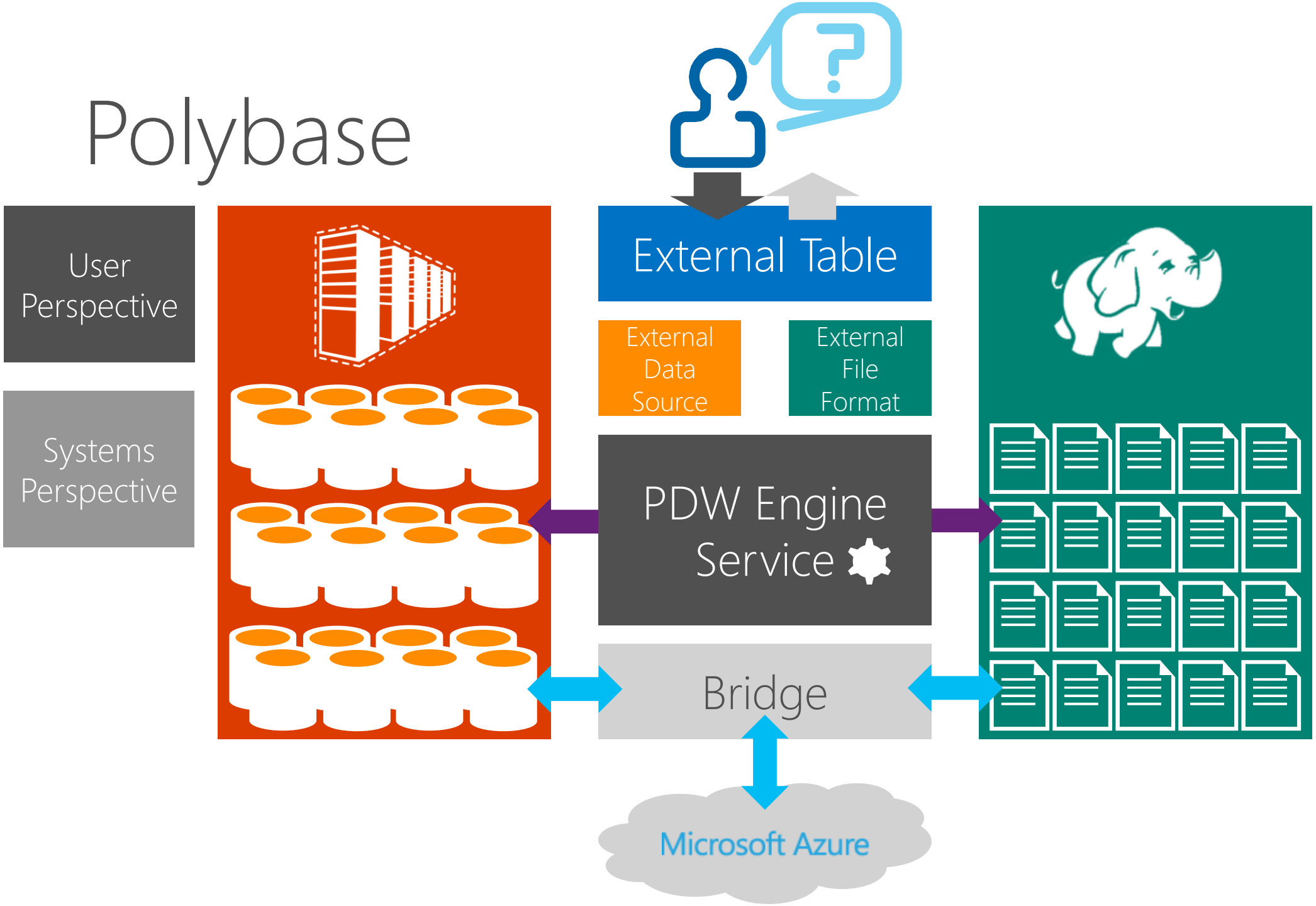
DMS shrink wraps HDFS Bridge with new "external" movement types

Bridge



*Control or Compute Node

Polybase


User Perspective
Systems Perspective
External Table
External Data Source
External File Format
PDW Engine Service
Bridge
Microsoft Azure

Balanced Execution

Essentially a divide and conquer challenge

- Break the task up into small enough pieces
- Spread those pieces round as evenly as we can

How do we do this for Hadoop?

Table Level Statistics

When an external table is created table level statistics are also persisted

- Row count
- Page count

Table statistics values

Row count

- 1000 rows
- Fixed default

Page count

- Based on file size as understood by Hadoop name node
- Converted to pages
- Influenced by compression

What are table statistics good for?

File Binding

- Verifies existence of file/folder
- Estimate row length & number of rows
- Sizes the file

Split Generation

- Calculate # of “splits” to allocate per compute node

What is a File Split?

- Fragment of an HDFS file
- <= HDFS file block size
  - Hadoop Region 256MB
  - Hadoop default 64MB
- The split is composed of two parts
  - an offset (in bytes) into the file
  - # of bytes to be read

Working Example of File Splitting

Example:

- 6 Node Appliance
- 720 GB File in Hadoop

Goal

- Each compute node to read an even share of the data

During MPP plan generation PDW calculates the file splits

Compute Nodes in Appliance	= 6
Reader/Writer threads per node	= 8
Source File in HDFS	= 720GB or 773,094,113,280 Bytes
Hadoop Region Block Size	= 256MB or 268,435,456 Bytes
Total # of File Splits (Volume / Block Size)	= 2880
# Splits per Compute Node	= 480
# Splits per Worker	= 60

Data Export & Data Movement

Exporting data with CETAS

CETAS – CREATE EXTERNAL TABLE AS SELECT

Post export three statements will be true

1. External table will now exist
2. Data will have been exported
3. Row & page count updated on external table

CETAS: Additional guidance

- Integration point is the file system
  - HDFS or WASB[s]
  - Not Hive or HCatalog
- Target is either a folder or a file
- Target does not have to already exist
- External table name must not exist in PDW DB
- Round-Tripping is perfectly possible
- Polybase will make a one-time best effort at clean-up

Export Data Movement Types

Three new "external" data movement types

- ExternalExportDistributedMove
  - Export DMS movement for distributed data in PDW
- ExternalExportReplicatedMove
  - Export DMS movement for replicated data in PDW
- ExternalExportControlMove
  - Export DMS movement for data that has already been "Master Moved" to the Control Node in PDW

ExternalExportDistributedMove

```
CREATE EXTERNAL TABLE HDFS_Web_Sales
WITH
(
     LOCATION     = '/TPCDS/web_sales/'
,    DATA_SOURCE = HDI_HadoopRegion_DataSource
,    FILE_FORMAT = HDI_HadoopRegion_RCFileLazyNoCompress
)
AS
SELECT ws.*
FROM   dbo.web_sales ws
JOIN   dbo.date_dim dd ON ws.ws_sold_date_sk = dd.d_date_sk
WHERE  dd.d_current_month = 'Y'
;
```

Explain: ExternalExportDistributedMove

```
<?xml version="1.0" encoding="utf-8"?>
<dsql_query>
  <sql>CREATE EXTERNAL TABLE HDFS_Web_Sales
WITH
(
     LOCATION    = '/TPCDS/web_sales/'
,    DATA_SOURCE = HDI_HadoopRegion_DataSource
,    FILE_FORMAT = HDI_HadoopRegion_RCFileLazyNoCompress
)
AS
SELECT ws.*
FROM
JOIN                    ON ws.ws_sold_date_sk = dd.d_date_sk
WHERE                   th = 'Y'</sql>
                        ost="0" total_number_operations="2">
    <                   ration_type="ExternalExportDistributedMove">
      <op               st cost="0" accumulative_cost="0" average_rowsize="0" outp
      <source           atement>...</source_statement>
      <external_uri>hdfs://HTUKIA-C-HHN01:8020/TPCDS/web_sales/</external_uri>
      <destination_table>[HDFS_Web_Sales]</destination_table>
    </dsql_operation>
  </dsql_operations>
  <meta-data>
    <full />
  </meta-data>
</dsql_query>
```

External Data Source

Movement Type

Target

Achieving Parallel Writes

- Exported Files use unique naming convention
  - {QueryID}_{YearMonthDay}_{HourMinutesSeconds}_{FileIndex}.txt
- Also good for lineage (presence of QID)
- File Index is zero based
  - Relationship is 1:1 with distributions
- File extension used depends on file format chosen
  - Txt
  - rcf

If the External Table is dropped and the same CETAS is re-executed then the target folder will have doubled its contents!

Parallel Writes for a Distributed Table

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
QID2077_20140402_233924_0.rcf	file	280.5 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_1.rcf	file	280.14 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_10.rcf	file	280.09 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_11.rcf	file	280.31 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_12.rcf	file	280.32 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_13.rcf	file	280.32 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_14.rcf	file	280.13 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_15.rcf	file	280.16 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_2.rcf	file	280.43 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_3.rcf	file	280.07 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_4.rcf	file	280.21 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_5.rcf	file	280.41 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_6.rcf	file	280.1 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_7.rcf	file	280.04 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_8.rcf	file	280.15 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup
QID2077_20140402_233924_9.rcf	file	280.67 MB	3	256 MB	2014-04-02 23:40	rw-r--r--	pdw_user	supergroup

ExternalExportReplicatedMove

```
CREATE EXTERNAL TABLE HDFS_Date_Dim
WITH
(
     LOCATION      = '/TPCDS/date_dim/'
,    DATA_SOURCE = HDI_HadoopRegion_DataSource
,    FILE_FORMAT = HDI_HadoopRegion_RCFileLazyNoCompress
)
AS
SELECT *
FROM   dbo.date_dim
;
```

Explain: ExternalExportReplicatedMove

```
<?xml version="1.0" encoding="utf-8"?>
<dsql_query>
  <sql>CREATE EXTERNAL TABLE HDFS_Date_Dim
WITH
(
     LOCATION     = '/TPCDS/date_dim/'
,    DATA_SOURCE = HDI_HadoopRegion_DataSource
,    FILE_FORMAT = HDI_HadoopRegion_RCFileLazyNoCompress
)
AS
SELECT
FROM                                  /sql>
  <dsq                                tal_cost="0" total_number_operations="2">
    <                                 peration_type="ExternalExportReplicatedMove">
      <op                             cost cost="0" accumulative_cost="0" average_rowsize=                    />
      <sourc                          statement>...</source_statement>
      <external_uri>hdfs://HTUKIA-C-HHN01:8020/TPCDS/date_dim/</external_uri>
      <destination_table>[HDFS_Date_Dim]</destination_table>
    </dsql_operation>
  </dsql_operati
  <meta-data>
    <full />
  </meta-data>
</dsql_query>
```

External Data Source

Movement Type

Target

Parallel Writes with Replicated Tables

- Are not attainable in the current version
- Replicated tables are written to a single file
- Only one replicated table will be queried

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
QID2063_20140402_232444_0.rcf	file	0.06 KB	3	256 MB	2014-04-02 23:24	rw-r--r--	pdw_user	supergroup

ExternalExportControlMove

```
CREATE EXTERNAL TABLE HDFS_Top10_Products
WITH
(
     LOCATION    = '/TPCDS/Top10_Products/'
,    DATA_SOURCE = HDI_HadoopRegion_DataSource
,    FILE_FORMAT = HDI_HadoopRegion_RCFileLazyNoCompress
)
AS
SELECT TOP (10)
       i_item_id
,      ws_item_sk
,      SUM(ws_net_profit) NetProfitCurrentMonth
FROM   dbo.web_sales ws
JOIN   dbo.date_dim dd ON ws.ws_sold_date_sk = dd.d_date_sk
JOIN   dbo.item     i  ON ws.ws_item_sk      = i.i_item_sk
WHERE  dd.d_current_month = 'Y'
GROUP BY
       i_item_id
,      ws_item_sk
;
```

Explain: ExternalExportControlMove

```
<?xml version="1.0" encoding="utf
<dsql_query>
  <sql>...</sql>
  <dsql_operations total_cost="0                    r_operations="6">
    <dsql_operation operation_type=                .</dsql_operation>
    <dsql_operation operation_type="ON              </dsql_operation>
    <dsql_operation operation_type="PARTITION_MOVE">
      <operation_cost cost="0.0696" accumulative_cost="0.0696" average_rowsize="29" output_rows="10" />
      <location distribution="AllDistributions" />
      <source_statement>...</source_statement>
      <destination>Control</destination>
      <destination_table>[TEMP_ID_12]</destination_table>
    </                ation>
    <                operation_type="ExternalExportControlMove">
                  ost cost="0" accumulative_cost="0.0696" average_rowsi                  " />
                  ement>...</source_statement>
      <external_uri>hdfs://HTUKIA-C-HHN01:8020/TPCDS/date_dim/</external_uri
      <destination t      >[HDFS_Top10_Products]</destination_table>
    </dsql_oper
    <dsql_ope                  type="ON">...</                  >
  </dsql_opera
  <meta-data>
    <full />
  </meta-data>
</dsql_query>
```

Initial Move to Control

Partition Destination

Movement Type

External Data Source

Target

Hybrid Queries

What are hybrid queries?

Read data from multiple external data sources

- HDFS
- PDW
- WASB[S]

Hybrid
=
Multitude of data sources accessed in a single query

External Data Movement Types

Three basic moves mirroring internal movement

- ExternalRoundRobinMove
- ExternalShuffleMove
- ExternalBroadcastMove

ExternalRoundRobinMove

```
SELECT *
FROM dbo.HDFS_Web_Sales
```

- Also known as the Random Hash
- Buffers re-distributed evenly across the compute nodes

Explain ExternalRoundRobinMove

```
<?xml version="  utf-8"?>
<dsql_query>
  <sql>...</sq
  <dsql_operat  3.798" total_number_operations="5">
    <dsql_operatio  on_type="RND_ID">...</dsql_operation>
    <dsql_operation o  ation_type="ON">
      <location permanent="false" distribution="AllDistributions" />
      <sql_operations>...</sql_operations>
    /dsql_operation>
    ql_operation operation_type="ExternalRoundRobinMove">
      peration_cost cost="3.798" accumulative_cost="3.798" av  output_rows="500" />
      ternal_uri>hdfs://HTUKIA-C-HHN01:8020/TPCDS/web_sales/</  ernal_uri>
      estination_table>[TEMP_ID_17]</destination_table>
    sql_operation>
    sql_operation operation_type="RETURN">
      <location distribut  "AllDistributions"
      <select>...</sel
    </dsql_operatio
    <dsql_operation  "ON">
      <location perm  istribution="AllDistributions" />
      <sql_operatio  ations>
    </dsql_operatio
  </dsql_operations>
</dsql_query>
```

Create Distributed Q Table

External Data Source

Movement Type

Target Q Table

Read From Q and Return

ExternalBroadcastMove

```
SELECT      i_item_id
,           s_store_id
FROM        dbo.HDFS_Item
CROSS JOIN  dbo.HDFS_Store
;
```

Both tables are external to PDW

- An external broadcast move is used as it is cheaper to broadcast immediately than it is to import the data and then broadcast

Explain: ExternalBroadcastMove

```
<?xml version="1.0" encoding="utf-8"?>
<dsql_query>
  <sql>...</sql>
  <dsql_operations total_cost="8.067" total_number_operations="9">
    <dsql_operation operation_type="RND_ID">
      <identifier>TEMP_ID_40</identifier>
    </dsq[illegible]ration>
    [illegible] operation_type="ON">
      [illegible]t="false" distribution="AllDistributions" />
      [illegible]...</sql_operations>
      [illegible]peration_type="ExternalRoundRobinMove">
        <operation_cost cost="2.445" accumulative_cost="2.445" ave[illegible]s="500" />
        <external_uri>hdfs://HTUKIA-C-HHN01:8020/TPCDS/store/</exter[illegible]
        <destination_table>[TEMP_ID_40]</destination_t[illegible]
    </dsql_operation>
    <dsql_operation operation_type="RND_ID[illegible]
      <identifier>TEMP_ID_41</identifier>
    </dsq[illegible]ration>
    [illegible] operation_type="ON">
      [illegible]t="false" distribution="AllComputeNodes" />
      [illegible]...</sql_operations>
      [illegible]peration_type="ExternalBroadcastMove">
        <operation_cost cost="5.622" accumulative_cost="8.067" ave[illegible]s="1000" />
        <external_uri>hdfs://HTUKIA-C-HHN01:8020/TPCDS/item/</externa[illegible]
        <destination_table>[TEMP_ID_41]</destination_t[illegible]
    </dsql_operation>
    <dsql_operation operation_type="RETURN[illegible]
    <dsql_operation operation_type="ON">...<[illegible]
    <dsql_operation operation_type="ON">...<[illegible]
  </dsql_operations>
</dsql_query>
```

1st Target

1st Movement

1st Source

2nd Target

2nd Movement

2nd Source

ExternalShuffleMove

```
SELECT
       i_item_id
,      ws_item_sk
,      SUM(ws_net_profit) NetProfitCurrentMonth
FROM   dbo.HDFS_web_sales ws
JOIN   dbo.date_dim dd ON ws.ws_sold_date_sk = dd.d_date_sk
JOIN   dbo.item     i  ON ws.ws_item_sk      = i.i_item_sk
WHERE  dd.d_current_month = 'Y'
GROUP BY
       i_item_id
,      ws_item_sk
OPTION (LABEL = 'External Shuffle Move')
;
```

Hybrid Query

Explain: ExternalShuffleMove

```
<?xml version="1.0" encoding="utf-8"?>
<dsql_query>
  <sql>...</sql>
  <dsql_operations total_cost="0.905" total_number_operations="5">
    <dsql_operation operation_type="RND_ID">
      <identifier>TEMP_ID_20</identifier>
    </dsql_operation>
    <dsql_operation operation_type="ON">
      rmanent="false" distribution="AllDistributions" />
      <sql_operations>...</sql_operations>
    </dsql_operation>
    <dsql_operation operation_type="ExternalShuffleMove">
      <operation_cost cost="0.905" accumulative_cost="0.905" 17" output_rows="500" />
      <external_uri>hdfs://HTUKIA-C-HHN01:8020/TPCDS/web_sales/ external_uri>
      <destination_table>[TEMP_ID_20]</destination_table>
      <shuffle_columns>ws_item_sk;</shuffle
    </dsql_oper
    <dsql ation_type="RETURN"> ation>
    <dsql_ ation_type="ON">...</dsql_operation>
  </dsql_op
</dsql_query>
```

Target

Movement Type

Source

Source

Data Import & Data Movement

Return of CTAS

Use CTAS to

- Perform a parallel import of data via Polybase
- Movement types are the same as hybrid

Additional steps included in the MPP plan

- Persist the results in PDW
- Check permissions
- Create extended properties
- Update Table level Statistics

Importing data with CTAS

```
CREATE TABLE Agg_ProductProfitCurrentMonth
WITH (DISTRIBUTION = HASH(ws_item_sk))
AS
SELECT
       i_item_id
,      ws_item_sk
,      SUM(ws_net_profit) NetProfitCurrentMonth
FROM   dbo.HDFS_web_sales ws
JOIN   dbo.date_dim dd ON ws.ws_sold_date_sk = dd.d_date_sk
JOIN   dbo.item     i  ON ws.ws_item_sk      = i.i_item_sk
WHERE  dd.d_current_month = 'Y'
GROUP BY
       i_item_id
,      ws_item_sk
OPTION(LABEL = 'CTAS : External Shuffle Move')
;
```

Console: ExternalShuffle

STEP ID	OPERATION	LOCATION	DISTRIBUTION	ROW COUNT
0	RandomIDOperation	Control	Unspecified	-1

```
TEMP_ID_67
```

1	OnOperation	Compute	AllDistributions	-1

```
CREATE TABLE [tempdb].[dbo].[TEMP_ID_67] ([ws_sold_date_sk] INT, [ws_item_sk] INT NOT NULL, [ws_net_profit] DECIM
```

2	HadoopShuffleOperation	DMS	Unspecified	-1

```
SELECT [T1_1].[ws_sold_date_sk] AS [ws_sold_date_sk],
       [T1_1].[ws_item_sk] AS [ws_item_sk],
       [T1_1].[ws_net_profit] AS [ws_net_profit]
FROM   [tpcds_cci].[dbo].[HDFS_Web_Sales] AS T1_1
```

3	ReturnOperation	Compute	AllDistributions	-1

```
SELECT [T1_1].[i_item_id] AS [i_item_id],
       [T1_1].[ws_item_sk] AS [ws_item_sk],
       [T1_1].[col] AS [col]
FROM   (SELECT   SUM([T2_2].[ws_net_profit]) AS [col],
                 [T2_1].[i_item_id] AS [i_item_id],
                 [T2_2].[ws_item_sk] AS [ws_item_sk]
        FROM     [tpcds_cci].[dbo].[item] AS T2_1
                 INNER JOIN
                 (SELECT [T3_2].[ws_item_sk] AS [ws_item_sk],
                         [T3_2].[ws_net_profit] AS [ws_net_profit]
                  FROM   (SELECT [T4_1].[d_date_sk] AS [d_date_sk]
                          FROM   [tpcds_cci].[dbo].[date_dim] AS T4_1
                          WHERE  ([T4_1].[d_current_month] = CAST (N'Y' COLLATE Latin1_General_100_CI_AS_KS_WS AS
                         INNER JOIN
                         [tempdb].[dbo].[TEMP_ID_67] AS T3_2
                         ON ([T3_1].[d_date_sk] = [T3_2].[ws_sold_date_sk])) AS T2_2
                 ON ([T2_1].[i_item_sk] = [T2_2].[ws_item_sk])
        GROUP BY [T2_1].[i_item_id], [T2_2].[ws_item_sk]) AS T1_1
```

4	OnOperation	Compute	AllDistributions	-1

```
DROP TABLE [tempdb].[dbo].[TEMP_ID_67]
```

- Create Q table
- External Shuffle into Q
- Return result read from Q
- Drop Q table

Console: ExternalShuffle /w CTAS

STEP ID	OPERATION	LOCATION	DISTRIBUTION	ROW COUNT
0	OnOperation	Control	Unspecified	-1

```
IF 0 = (SELECT HAS_PERMS_BY_NAME(N'[tpcds_cci]', 'DATABASE', 'CREATE TABLE', null, null) & HAS_PERMS_BY_NAME(N'[c
BEGIN
RAISERROR (N'6004;User does not have permission to perform this action.', 14, 9)
END
```

1	OnOperation	Compute	AllDistributions	-1

```
CREATE TABLE [tpcds_cci].[dbo].[Agg_ProductProfitCurrentMonth] ([i_item_id] CHAR(16) COLLATE Latin1_General_100_C
ON [PRIMARY] WITH(DATA_COMPRESSION=PAGE);
```

2	RandomIDOperation	Control	Unspecified	-1
3	OnOperation	Compute	AllDistributions	-1
4	HadoopShuffleOperation	DMS	Unspecified	-1
5	OnOperation	Compute	AllDistributions	80166

```
INSERT INTO [tpcds_cci].[dbo].[Agg_ProductProfitCurrentMonth] WITH (TABLOCK) ([i_item_id], [ws_item_sk], [NetProf
SELECT [T1_1].[i_item_id],
       [T1_1].[ws_item_sk],
       [T1_1].[col]
FROM   (SELECT   SUM([T2_2].[ws_net_profit]) AS [col],
                 [T2_1].[i_item_id] AS [i_item_id],
                 [T2_2].[ws_item_sk] AS [ws_item_sk]
        FROM     [tpcds_cci].[dbo].[item] AS T2_1
                 INNER JOIN
                 (SELECT [T3_2].[ws_item_sk] AS [ws_item_sk],
                         [T3_2].[ws_net_profit] AS [ws_net_profit]
                  FROM   (SELECT [T4_1].[d_date_sk] AS [d_date_sk]
                          FROM   [tpcds_cci].[dbo].[date_dim] AS T4_1
                          WHERE  ([T4_1].[d_current_month] = CAST (N'Y' COLLATE Latin1_General_100_CI_AS_KS_WS AS
                         INNER JOIN
                         [tempdb].[dbo].[TEMP_ID_85] AS T3_2
                         ON ([T3_1].[d_date_sk] = [T3_2].[ws_sold_date_sk])) AS T2_2
                 ON ([T2_1].[i_item_sk] = [T2_2].[ws_item_sk])
        GROUP BY [T2_1].[i_item_id], [T2_2].[ws_item_sk]) AS T1_1
OPTION (MAXDOP 1)
```

6	OnOperation	Compute	AllDistributions	-1

- Check permissions
- Create Table
- Create Q
- External Shuffle into Q
- Populate persistent table from Q
- Drop Q Table

Console: ExternalShuffle /w CTAS

STEP ID	OPERATION	LOCATION	DISTRIBUTION	ROW COUNT
0	OnOperation	Control	Unspecified	-1
1	OnOperation	Compute	AllDistributions	-1
2	RandomIDOperation	Control	Unspecified	-1
3	OnOperation	Compute	AllDistributions	-1
4	HadoopShuffleOperation	DMS	Unspecified	-1
5	OnOperation	Compute	AllDistributions	80166
6	OnOperation	Compute	AllDistributions	-1
DROP TABLE [tempdb].[dbo].[TEMP_ID_85]				
7	OnOperation	Control	Unspecified	-1
CREATE TABLE [tpcds_cci].[dbo].[Agg_ProductProfitCurrentMonth] ([i_item_id] CHAR(16) COLLATE Latin1_General_100_(ON [PRIMARY] WITH(DATA_COMPRESSION=PAGE);				
8	OnOperation	Control	Unspecified	-1
EXEC [tpcds_cci].[sys].[sp_addextendedproperty] @name=N'pdw_physical_name', @value=N'Table_07b71db03dc440a8a140f8 @level1name=N'Agg_ProductProfitCurrentMonth'				
9	OnOperation	Control	Unspecified	-1
EXEC [tpcds_cci].[sys].[sp_addextendedproperty] @name=N'pdw_distribution_type', @value=N'Distributed', @level0typ @level1name=N'Agg_ProductProfitCurrentMonth'				
10	OnOperation	Control	Unspecified	-1
EXEC [tpcds_cci].[sys].[sp_addextendedproperty] @name=N'pdw_distribution_column', @value=N'ws_item_sk', @level0ty @level1name=N'Agg_ProductProfitCurrentMonth'				
11	DbccShowStatisticsOperation	Compute	AllDistributions	-1
[tpcds_cci].sys.sp_executesql @statement=N'DBCC SHOW_STATISTICS ([Agg_ProductProfitCurrentMonth]) WITH STATS_STRI				
12	OnOperation	Control	Unspecified	-1
UPDATE STATISTICS [tpcds_cci].[dbo].[Agg_ProductProfitCurrentMonth] WITH ROWCOUNT = [ROWCOUNT_TEMP_ID_86], PAGEC(				

- Create Persisted Table
- Add Extended properties
- Update table level statistics

Explain: External Shuffle w/ CTAS 1/2

```
<?xml version="1.0" encodi
<dsql_query>
  <sql>...</sql>
  <dsql_operations total_c                    total_number_operations="6">
    <dsql_operation operat
      <location permanent="tr                  ution="AllDistributions" />
      <sql_operations>
        <sql_operation type="statement">CREATE TABLE [tpcds_cci].[dbo].[Agg_ProductProfitCurrentMonth] ([i_item_id
      </sql_operations>
    </dsql_operation>
    <dsql_operation operation_type="RND_ID">
      <identifier>TEMP_ID_92</identifier>
    </dsql_operation>
    <dsql_operation operation_type="ON">
      <location permanent="false" distribution="AllDi                    ns" />
                                  >
                                  n type="statement">CREATE TABLE [tempdb].[dbo].[TE MP_ID_92] ([ws_sold_date_sk] INT, [ws_item_s
                                  s>
                                  peration_type="ExternalShuffleMove">
      <op                         ost cost="140.946912" accumulative_cost="140.946                    17" output_rows="79362" />
      <exte                       _uri>hdfs://HTUKIA-C-HHN01:8020/TPCDS/web_sales/</ext                    l_uri>
      <destination_table>[TEMP_ID_92]</destinat                    able>
      <shuffle_columns>ws_item_sk;</shuffl
    </dsql_operation>
```

Final Target

Target Q (for Shuffle)

Target Q (for Shuffle)

Movement Type

Shuffle Key

External Data Source

Explain: External Shuffle w/ CTAS 2/2


```xml
<?xml version="1.0" encoding="utf-8"?>
<dsql_query>
  <sql>...</sql>
  <dsql_operations total_cost="140.946912" total_number_operations="6">
    <dsql_operation operation_type="ON">...</dsql_operation>
    <dsql_operation operation_type="RND_ID">...</dsql_operation>
    <dsql_operation operation_type="ON">...</dsql_operation>
    <dsql_operation operation_type="ExternalShuffle">...</dsql_operation>
    <dsql_operation operation_type="ON">
      <location permanent="true" distribution="AllDistributions"
      <sql_operations>
        <sql_operation type="statement">...</sql_operation>
      </sql_operations>
    </dsql_operation>
    <dsql_operation operation_type="ON">
      <location permanent="false" distribution="AllDistributions" />
      <sql_operations>
        <sql_operation type="statement">DROP TABLE [tempdb].[dbo].[TEMP_ID_92]</sql_operation>
      </sql_operations>
    </dsql_operation>
  </dsql_operations>
  <meta-data>
    <partitioned>
      <partitioning_column index="2" />
    </partitioned>
  </meta-data>
</dsql_query>
```

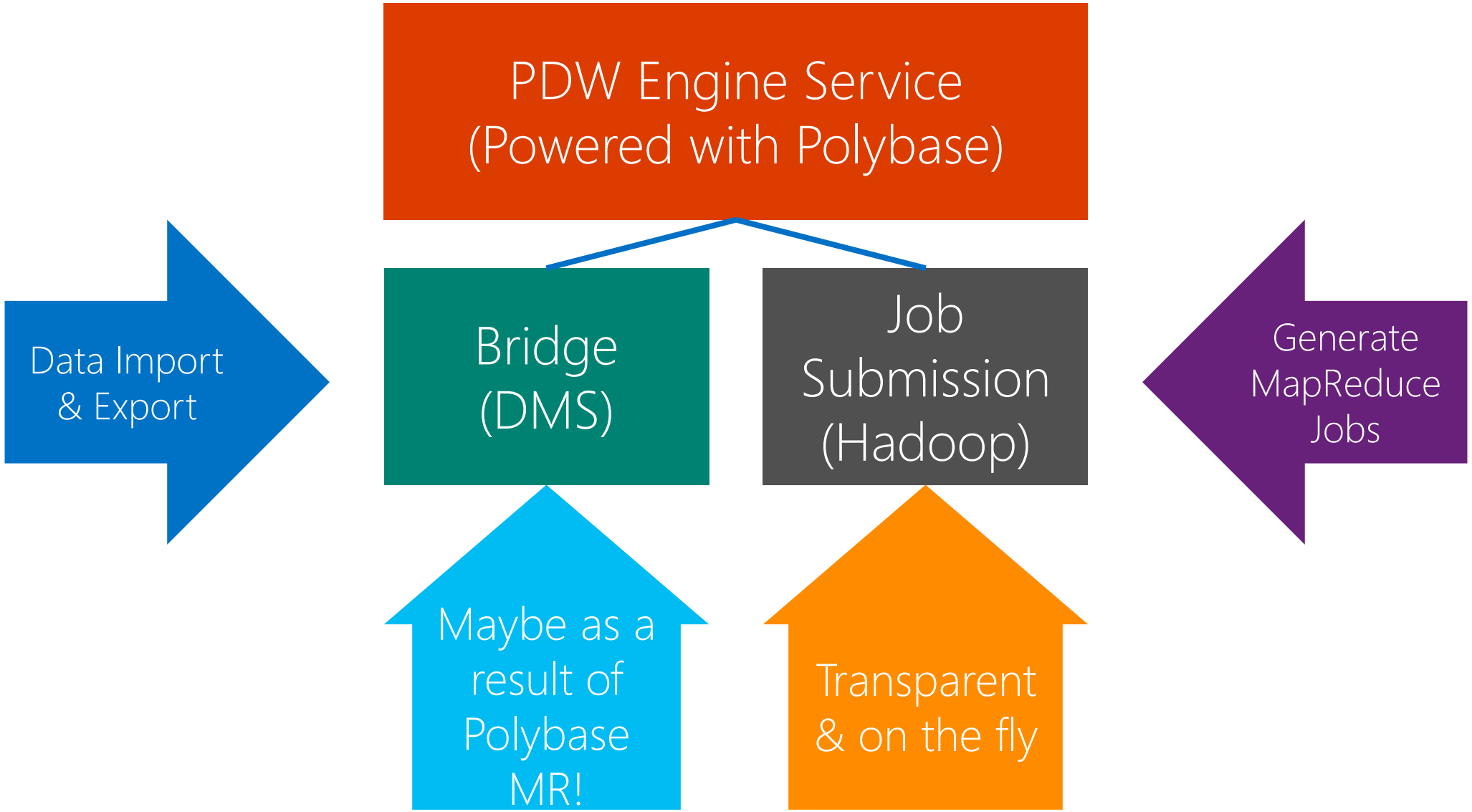

External Movement

Insert to Target (MAXDOP 1)

Drop Q

Split Query Execution

Split Query Processing



Using Split Query

Map Job designed to minimize movement

- Push predicates down to remote data store
- Reduce data volume to transfer

Understanding Overheads

- Table level stats only give size of table
- Selectivity of data needs to be considered
- Map job output must be persisted in Hadoop
- Need additional data to decide!

Column Level Statistics

Provides the additional data we need

- Crucial for cardinality estimation
- Enabled for External Tables
- Manual operation
- CREATE / DROP Only – not Update

Understanding Costs

- Submitting Hadoop jobs is costly
- Spin-up time ~20-30 seconds

Consequently…

- If PDW Engine estimates (based on stats) an execution time of less than 20-30 seconds there will be no push down

Pushdown trigger point

Push down will <u>not</u> be considered for:

- Data Transfers < 1GB per distribution
- Faster to simply import the data

Map Job

- Scan Columns
  - Simple projection
- Filter Rows
  - Push-able expressions
- Project Columns
  - Calculate expressions
- Materialize data
  - Persist data in temporary output directory

Pushdown Predicate Example

Example:

Compute Nodes in PDW	= 6
Distributions in total	= 48
Data to be transferred > 48GB	= Pushdown
Data to be transferred <= 48GB	= No Pushdown

Scoped Functionality

In

- Selection
  - Filter rows
- Projection
  - Filter columns

Out

- Push down JOINs
- Aggregation
  - Partial aggregation
  - Final aggregation

Selection: filter rows

```
SELECT    *
FROM      HDFS_Customer c
WHERE     c.account_balance < 20000

SELECT    *
FROM      HDFS_Customer c
WHERE     c.JobTitle IN ('Developer', 'Tester')

SELECT    *
FROM      HDFS_Clickstream c
WHERE     c.URI = 'www.microsoft.com'
AND       c.IP_address BETWEEN 127.0.0.1 AND 127.0.0.7
```

Projection: filter columns

```
SELECT    c.name, c.first_name+' '+c.last_name
FROM      HDFS_Customer c
WHERE     c.account_balance < 20000

SELECT    c.ac
FROM      HDFS_Customer c
WHERE     c.JobTitle IN ('Developer', 'Tester')

SELECT    c.click
FROM      HDFS_Clickstream c
WHERE     c.URI = 'www.microsoft.com'
AND       c.IP_address BETWEEN 127.0.0.1 AND 127.0.0.7
```

Supported Operators

Comparison Operators: Decimal & Datetime

<

>

=

!= <>

>=

<=

Arithmetic Operators: Decimal

+

-

*

/

%

Supported Operators

Logical Operators

AND

OR

NOT

IS NULL

IS NOT NULL

Unary Operators

+()

-()

~()

“It Depends” Operators

- BETWEEN
- LIKE
- NOT
- IN

It depends because…
The query optimizer may re-write the operator in a way that does not support pushdown

That said…
The query optimizer tends to re-write these operators as primitive relational operators which can be pushed down

Partial Pushdown

What happens when a table has pushable and non-pushable predicates?

- pushable predicates sent to Map job
- Non-pushable filters applied in PDW

External Pushdown operations

Hadoop Operation

- Represents the query sub-tree executed via MapReduce to support predicate pushdown

HadoopFileOperation

- File operation executed on Hadoop Cluster
- Delete temporary job files (post pushdown)

Configuration & Monitoring

sp_configure

Option Hadoop Connectivity	Value
Disable Hadoop Connectivity	0
Hortonworks for Windows Server (HDP 1.3)	1
HDInsight on Analytics Platform System (HDP 1.3)	1
HDInsight Windows Azure Blob Storage (WASB[S])	1
Hortonworks for Linux (HDP 1.3)	2
Cloudera CDH 4.3 for Linux	3

sp_configure

```
EXEC sp_configure;
EXEC sp_configure 'hadoop connectivity',1;
RECONFIGURE;
--Now Restart PDW Region
EXEC sp_configure;
```

	name	minimum	maximum	config_value	run_value
1	hadoop connectivity	0	3	1	0
2	redistribute mode	0	1	0	0

	name	minimum	maximum	config_value	run_value
1	hadoop connectivity	0	3	1	1
2	redistribute mode	0	1	0	0

You will need to stop and start the appliance after changing this value

Core-site.xml changes for WASB

To create WASB[s] external data sources

- Get storage account access key from Azure
- Add to core-site.xml file in PDW (control node)

APS Control Node

Core-site.xml (append)

```
<property>
    <name>fs.azure.account.key.<your storage account name>.blob.core.windows.net</name>
    <value><your storage account access key></value>
</property>
```

N.B. Any user with CONTROL SERVER or ALTER ANY EXTERNAL DATA SOURCE permission can create an external data source that accesses this account which in turn can be consumed by users with create table permission. EXTERNAL DATA SOURCEs exist at the PDW level

Hadoop Weak Authentication

Is weak authentication is enabled?

- dfs.permission = true
- Create a user PDW_User
- Grant PDW_User full read/write permissions
- All Polybase calls are made by this user with this security context

DMVs & Catalog Views

Catalog Views

- sys.external_tables
- sys.external_data_sources
- Sys.external_file_formats

DMV

- sys.dm_pdw_dms_external_work
- sys.dm_pdw_hadoop_operations

Summary

In this module you learned…

- What Polybase is
- Why it exists
- Why it is both important and innovative
- Polybase goals and how it achieves them
- How to configure Polybase for agnostic access