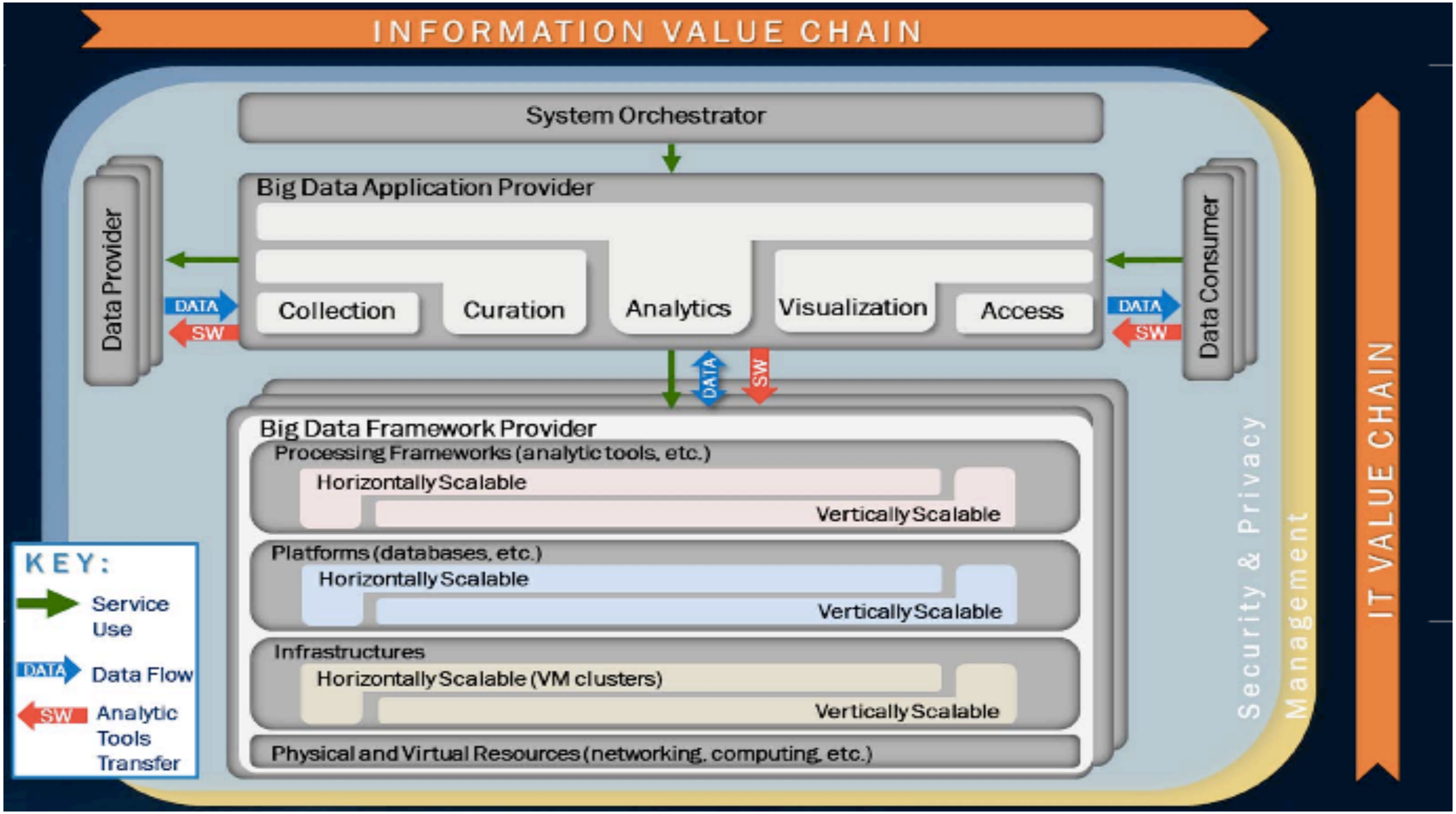
Generic Big Data Use Cases
+ Sample Mappings to Reference Architecture
+ Potential Standards
+ Implementation Examples

Bob Marcus
robert.marcus@et-strategies.com

NIST Reference Architecture



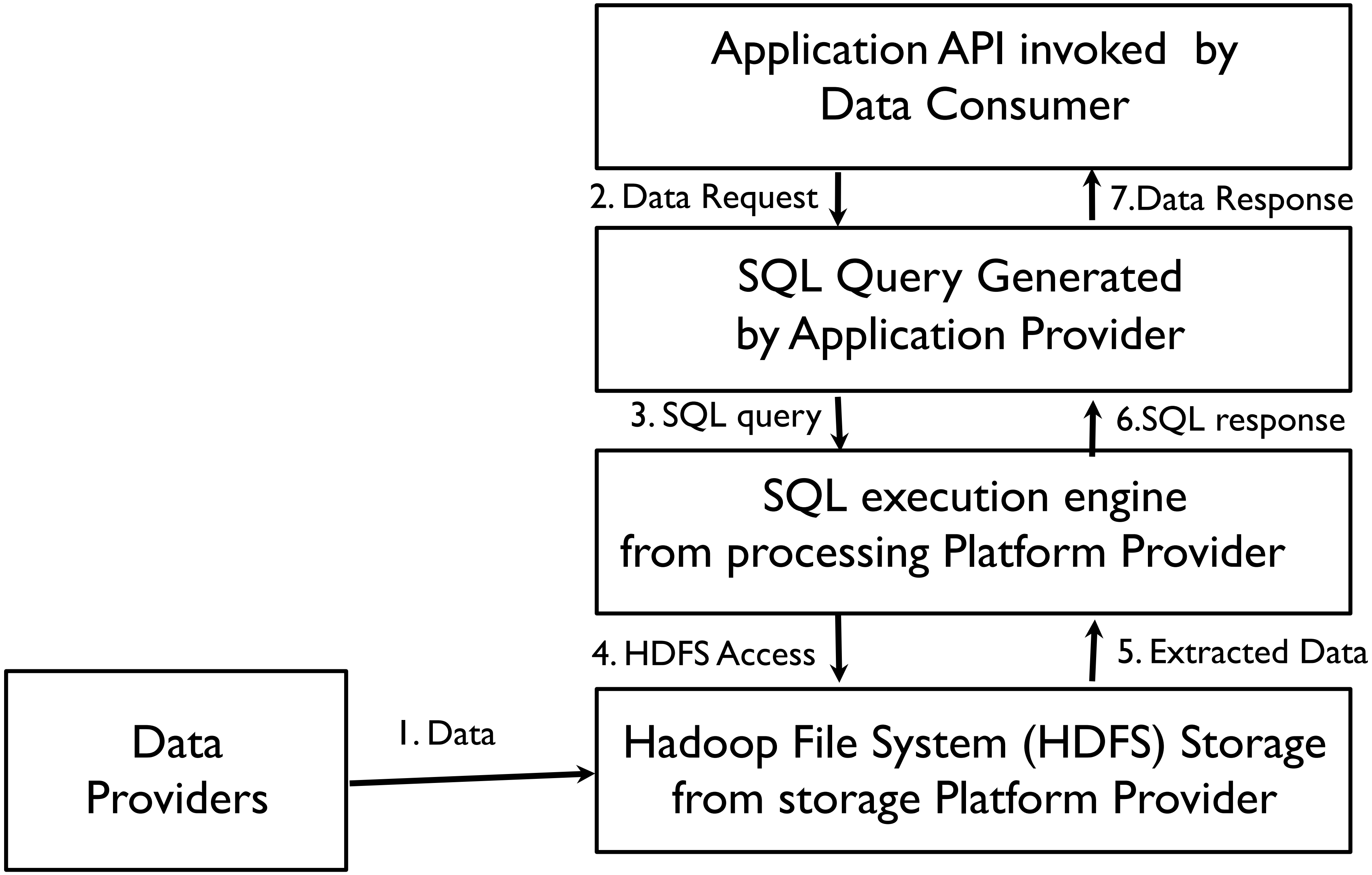
1. Generic Use Case

Use Case: Multiple users performing interactive queries and updates on databases with basic availability

Mapping:

- Data Providers and Consumers are users.
- Framework Providers manage the databases.
- Application Providers define schemas for the databases.
- Users interact with the databases on the Framework Provider using the schemas supplied by the Application Provider.

1. Example: Interactive SQL Query to HDFS Data



Application API invoked by Data Consumer

2. Data Request

7.Data Response

SQL Query Generated by Application Provider

3. SQL query

6.SQL response

SQL execution engine from processing Platform Provider

4. HDFS Access

5. Extracted Data

Data Providers

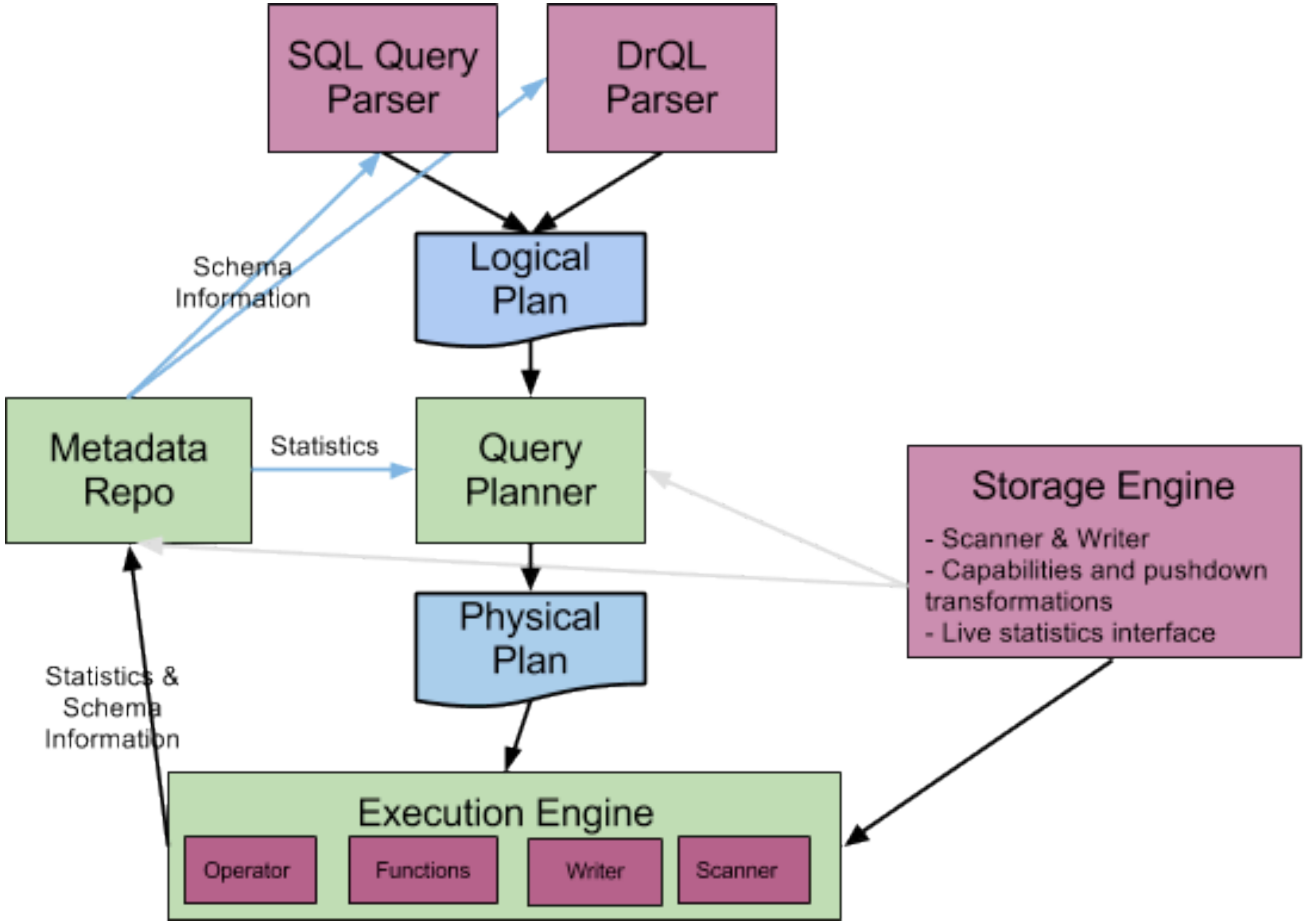
1. Data

Hadoop File System (HDFS) Storage from storage Platform Provider

Note that Data Provider sends the Raw Data directly to the Storage Platform Provider. This case can not be mapped directly to the Reference Architecture

Potential Standard: SQL-like language targeted at Horizontally Scalable Data Sources

1. SQL on Hadoop Example (Drill)



Apache Drill

Reference: http://incubator.apache.org/drill/

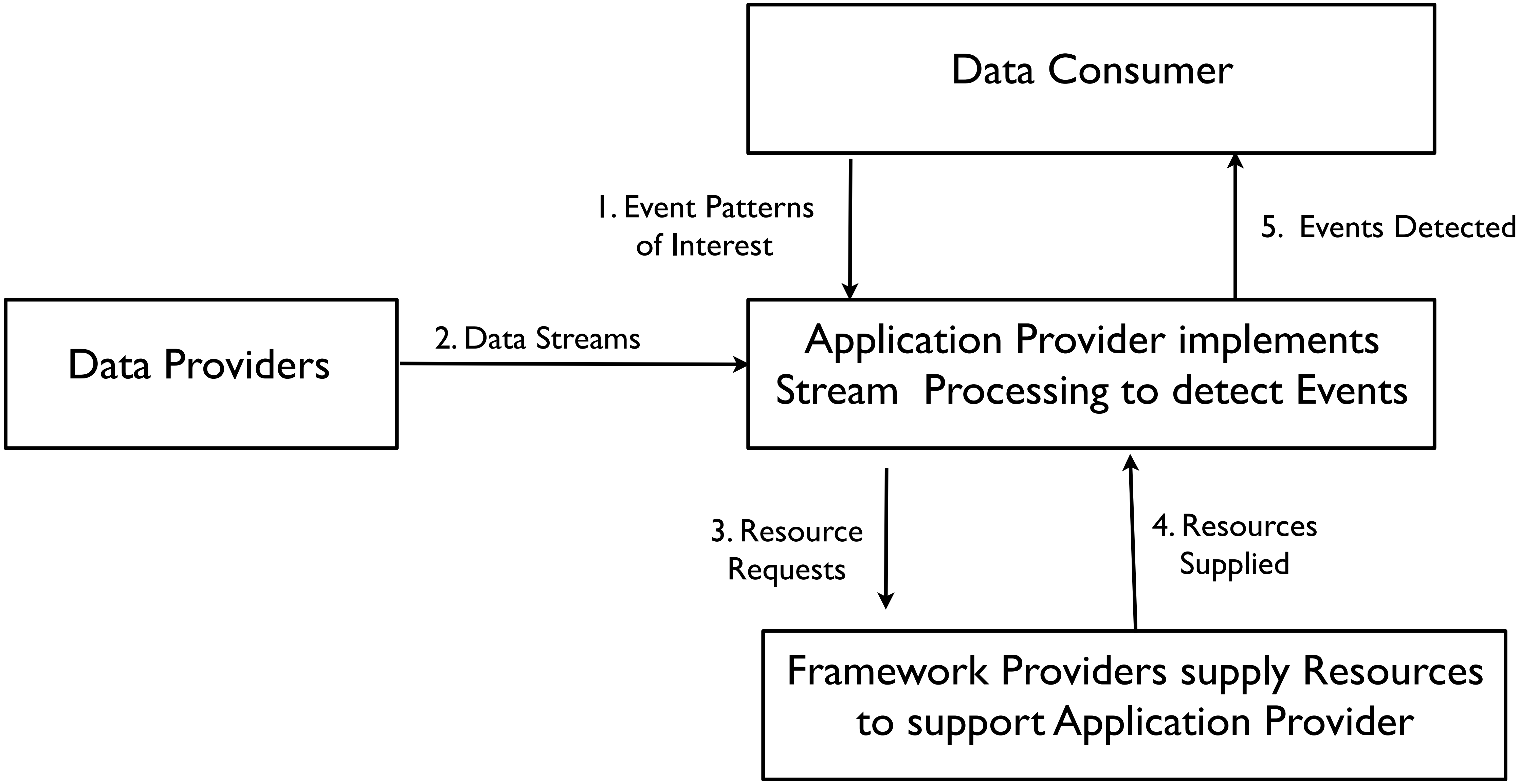
2. Generic Use Case

Use Case: Perform real-time analytics on data streams and send outputs to end-users and/or data stores

Mapping:

- Data Providers supply data streams.
- Application Provider supplies code to process data streams, detect events, and notify users.
- Framework Providers supplies resources to run the code.
- Data Consumers receive notification of events

2. Example: Event Notification to Consumer



Data Consumer

1. Event Patterns of Interest

5. Events Detected

Data Providers

2. Data Streams

Application Provider implements Stream Processing to detect Events

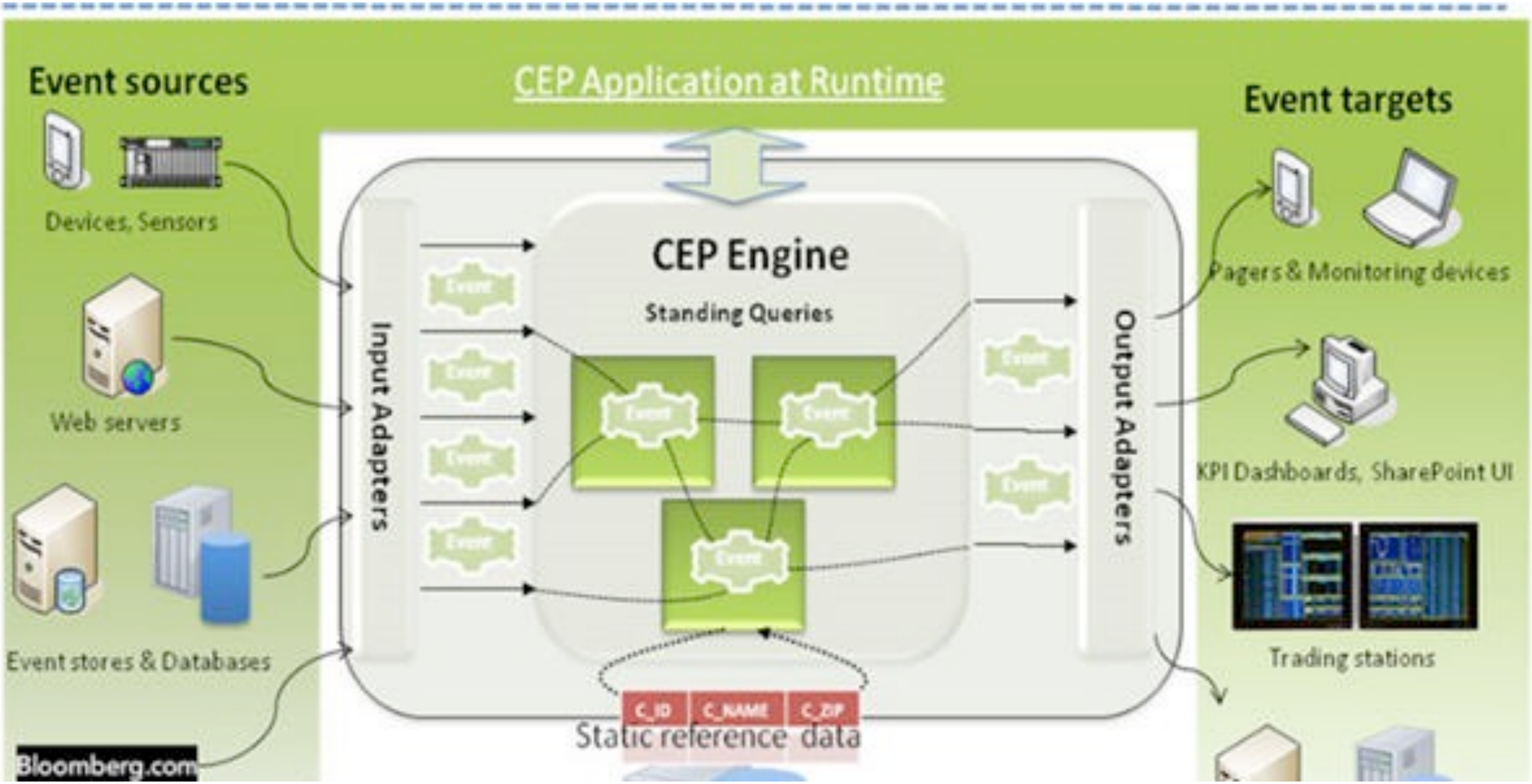
3. Resource Requests

4. Resources Supplied

Framework Providers supply Resources to support Application Provider

Potential Standard: Event Pattern Query Language and Event Description Language

2. Stream Processing Example (StreamInsight)



StreamInsight Event Processing Architecture from Microsoft

Reference: http://www.jasonvolpe.com/topics/data-warehousing/

Apache Drill

Reference: http://incubator.apache.org/drill/

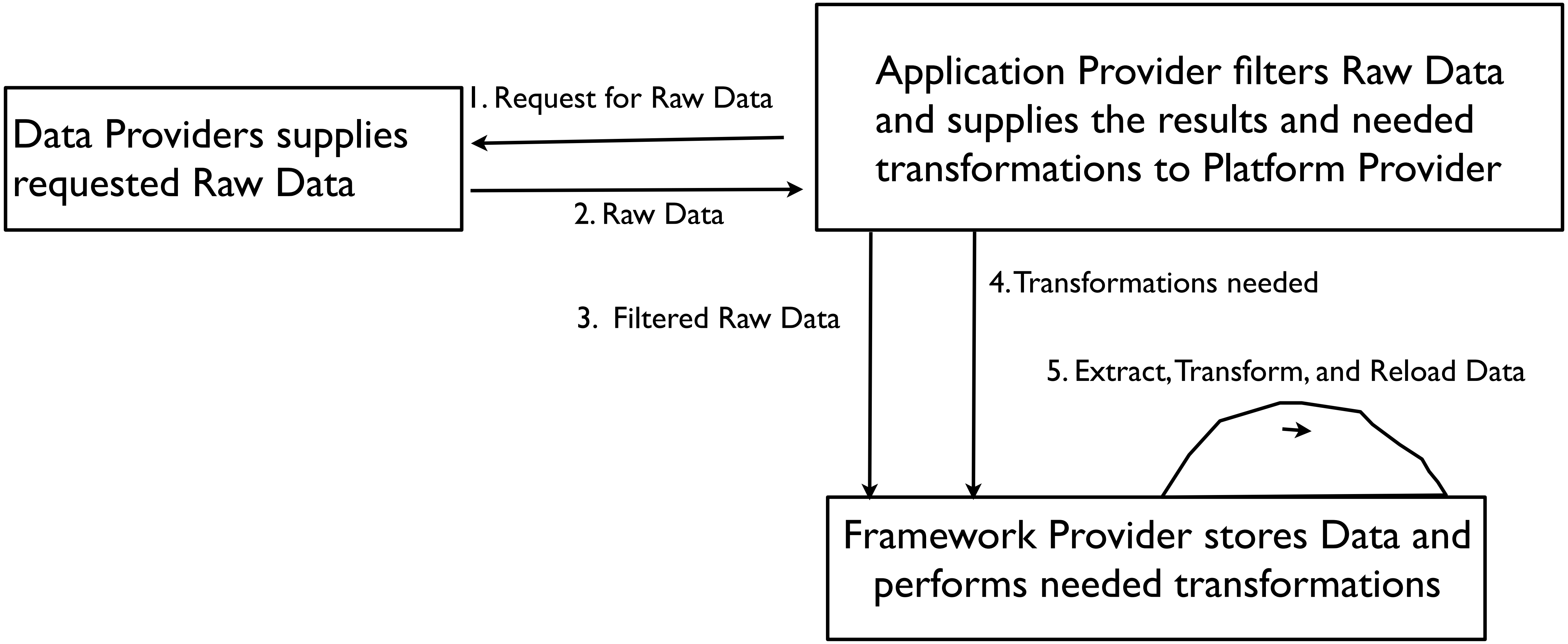
3. Generic Use Case

Use Case: Move data from external data sources into a horizontally scalable data store, transform it using horizontally scalable processing and return it to the horizontally scalable data store (ELT)

Mapping:

- Data Provider is the external data source.
- Framework Provider supplies horizontally scalable data store.
- Application Provider supplies transformation used by Framework Provider for horizontally scalable processing and returning results to data store.

3. Example: Raw Data Transformation (ELT)



Data Providers supplies requested Raw Data

1. Request for Raw Data

2. Raw Data

Application Provider filters Raw Data and supplies the results and needed transformations to Platform Provider

3. Filtered Raw Data

4. Transformations needed

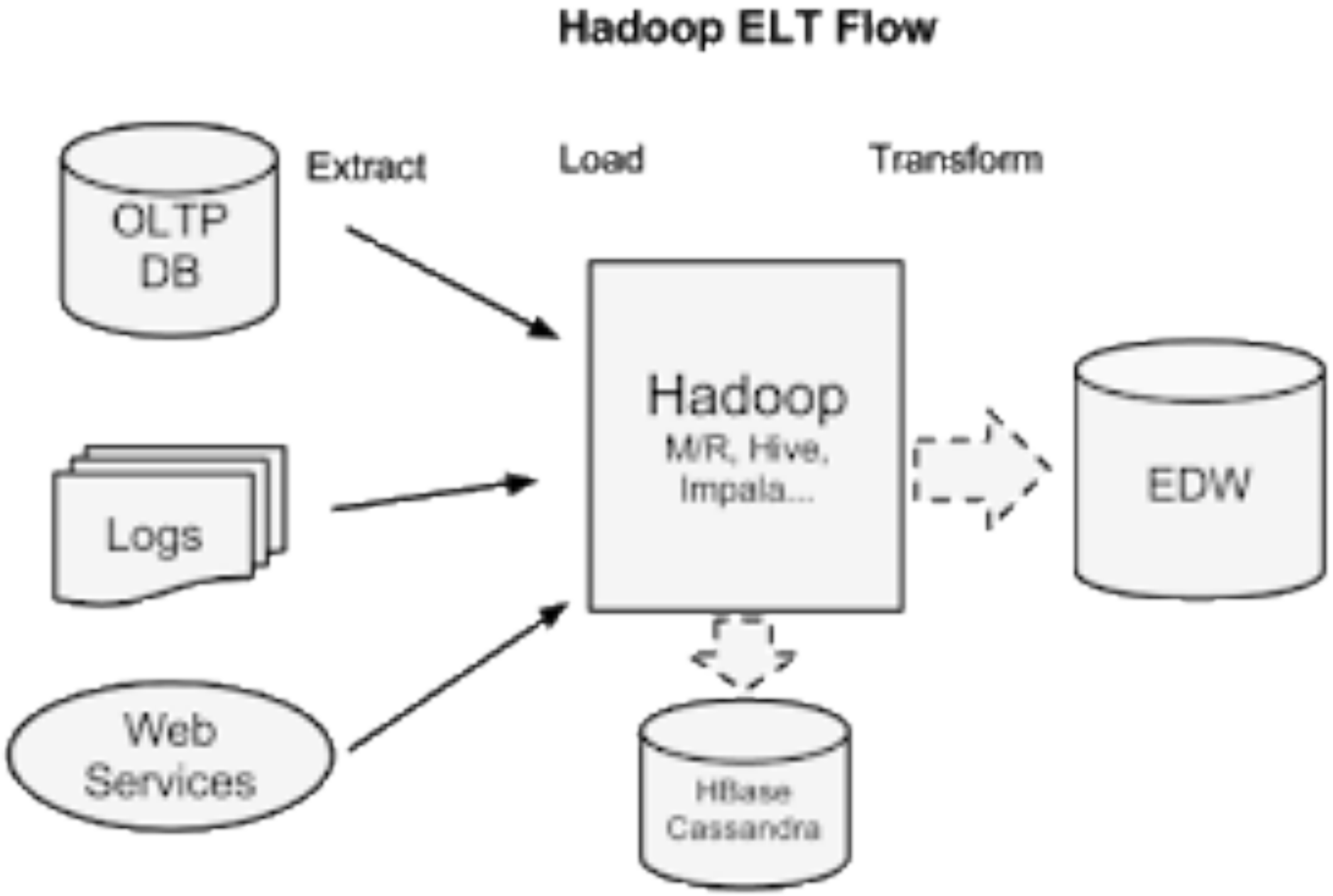
5. Extract, Transform, and Reload Data

Framework Provider stores Data and performs needed transformations

Note that there is a related Use Case where the Data Provider sends the Raw Data directly to the Storage Platform with no Application Provider Filtering. This Use Case can not be mapped to the current Reference Architecture

Potential Standards: Declarative Transformation DescriptionsFormats for Files, Data, and Metadata

3: Raw Data Transformation Example



Hadoop-Based Extract-Load-Transform (ELT)

Reference: http://www.dzone.com/articles/hadoop-t-etl

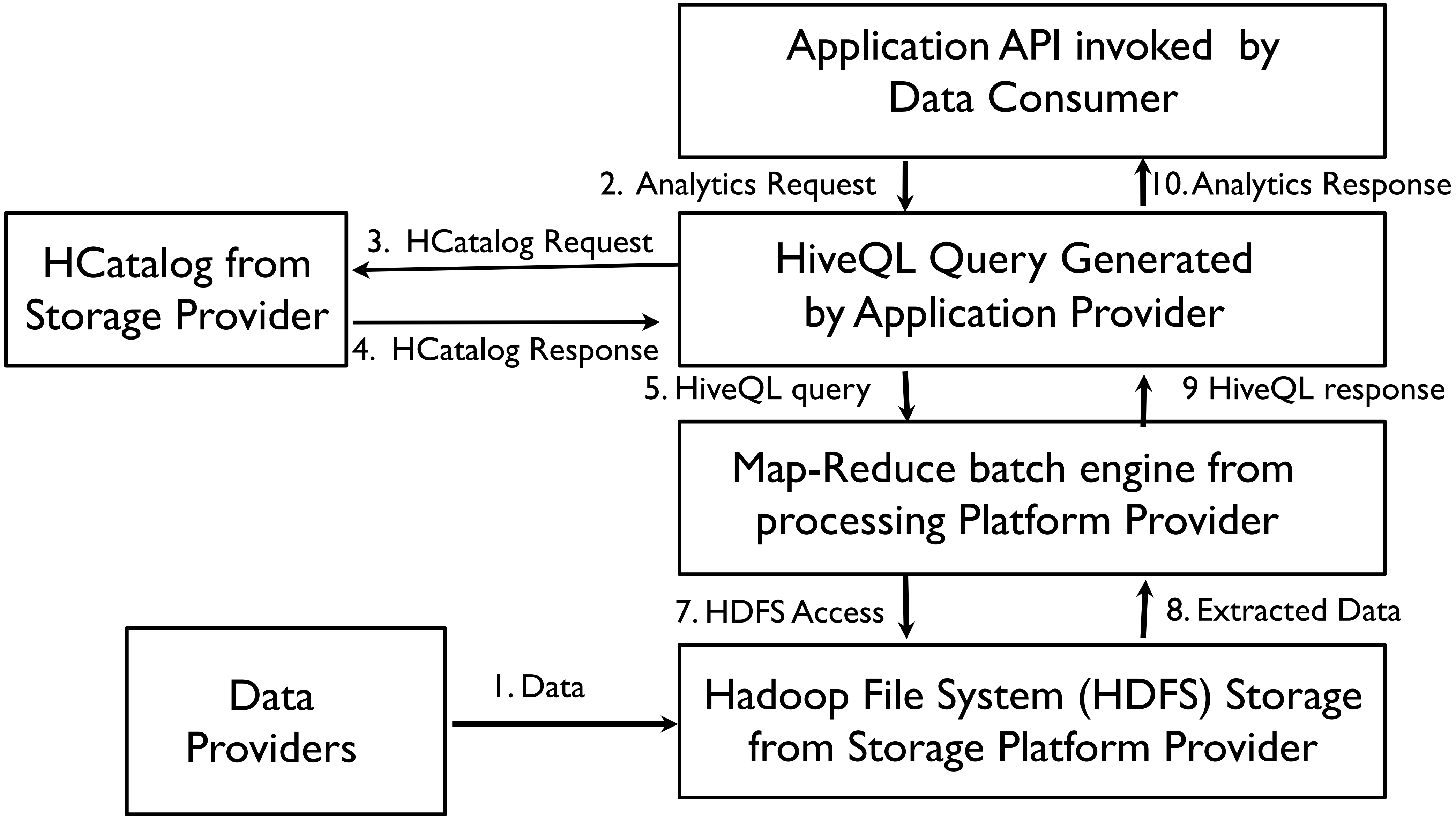
4. Generic Use Case

Use Case: Perform batch analytics on the data in a horizontally scalable data store using horizontally scalable processing with a user-friendly interface

Mapping:

- Data Providers supply data for the horizontal scalable data store
- Framework Provider supplies the horizontally scalable data store, metadata catalog, and processing engine.
- Application Provider supplies the analytic software that forms the basis of the processing.
- User selects the appropriate analytics to be performed.

4. Example: HiveQL Analytic Query to HDFS



Application API invoked by Data Consumer

2. Analytics Request

10. Analytics Response

HiveQL Query Generated by Application Provider

3. HCatalog Request

4. HCatalog Response

HCatalog from Storage Provider

5. HiveQL query

9 HiveQL response

Map-Reduce batch engine from processing Platform Provider

7. HDFS Access

8. Extracted Data

Hadoop File System (HDFS) Storage from Storage Platform Provider

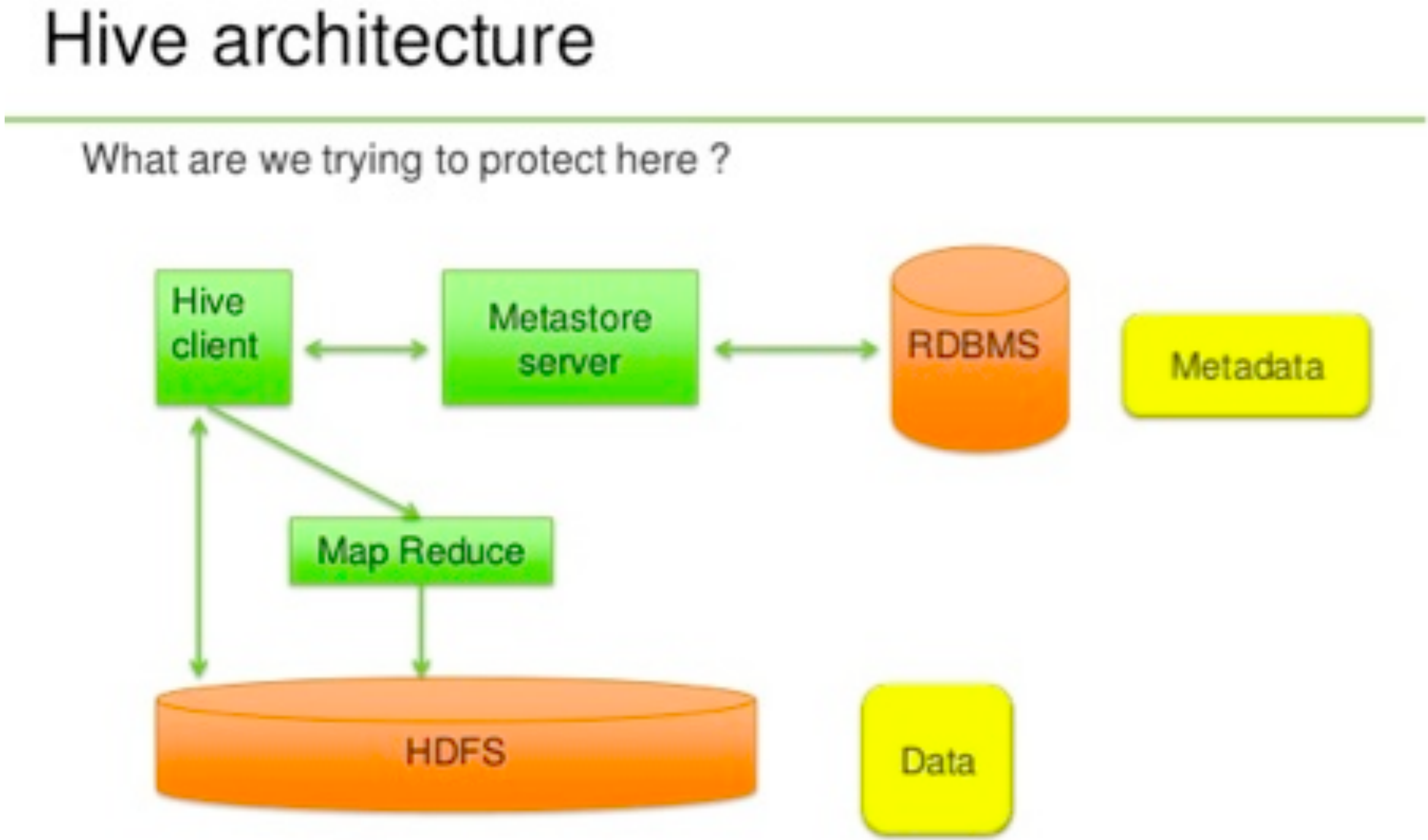
1. Data

Data Providers

Note that Data Provider sends the Data directly to the Storage Platform Provider. This case can not be mapped directly to the Reference Architecture

Potential Standards: SQL-like batch language (e.g HiveQL) targeted at Horizontally Scalable Data Sources
Metadata Catalog interface (e.g. HCatalog) to support diverse interfaces (e.g. Pig)

4. Example: HiveQL Analytic Query to HDFS



Hive architecture

What are we trying to protect here ?

Hive Architecture and Authorization

Reference: http://venublog.com/2013/07/16/hadoop-summit-2013-hive-authorization/

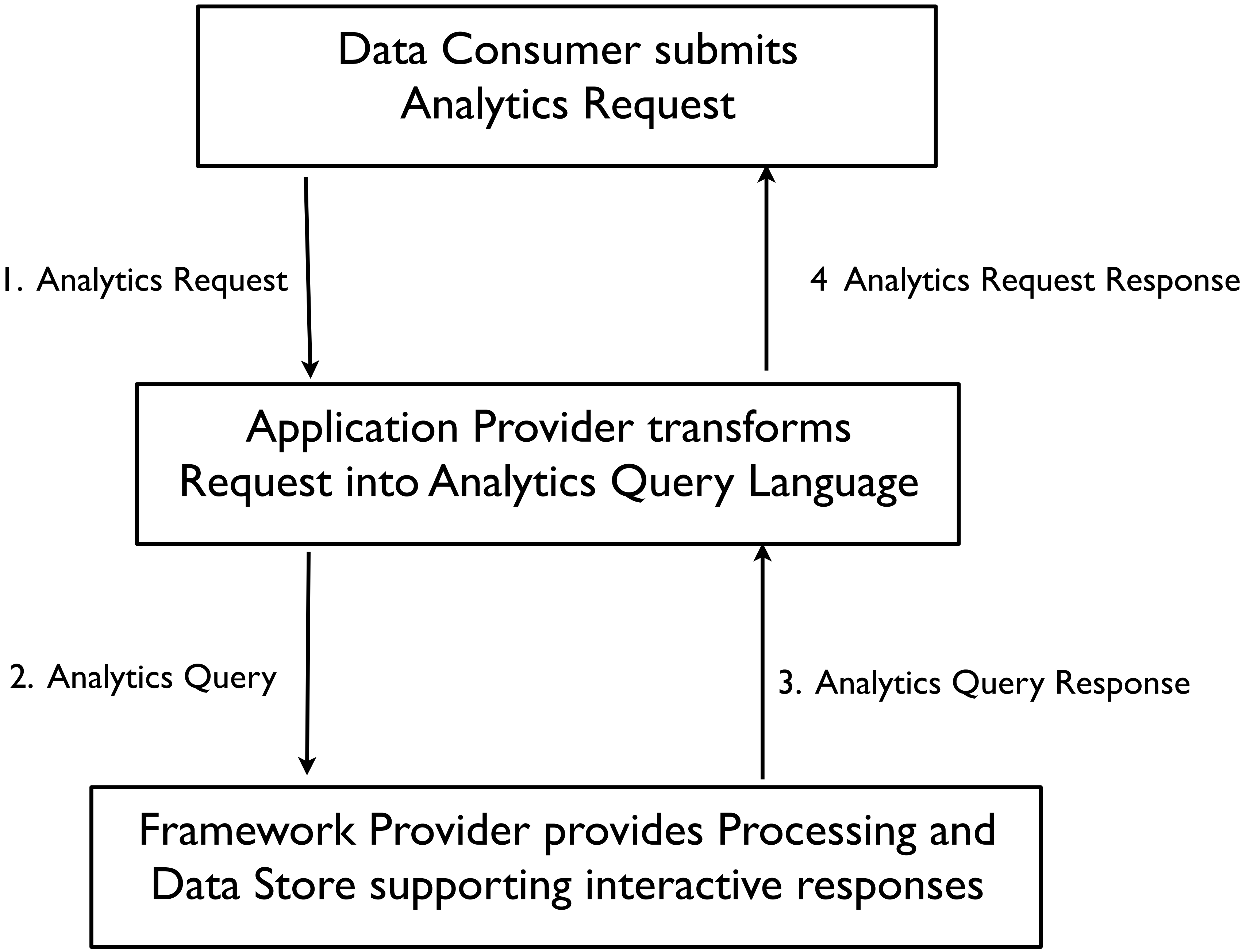
5. Generic Use Case

Use Case: Perform interactive analytics on data in an analytics-optimized database (e.g. in memory or Enterprise Data Warehouse)

Mapping:

- Framework Provider supplies the platform and basic database capabilities.
- Application Provider supplies optimized schema and configuration to support analytic queries.
- Data Consumer submits analytic queries.

5. Example: Interactive Analytics Request to Data



Data Consumer submits Analytics Request

1. Analytics Request

4 Analytics Request Response

Application Provider transforms Request into Analytics Query Language

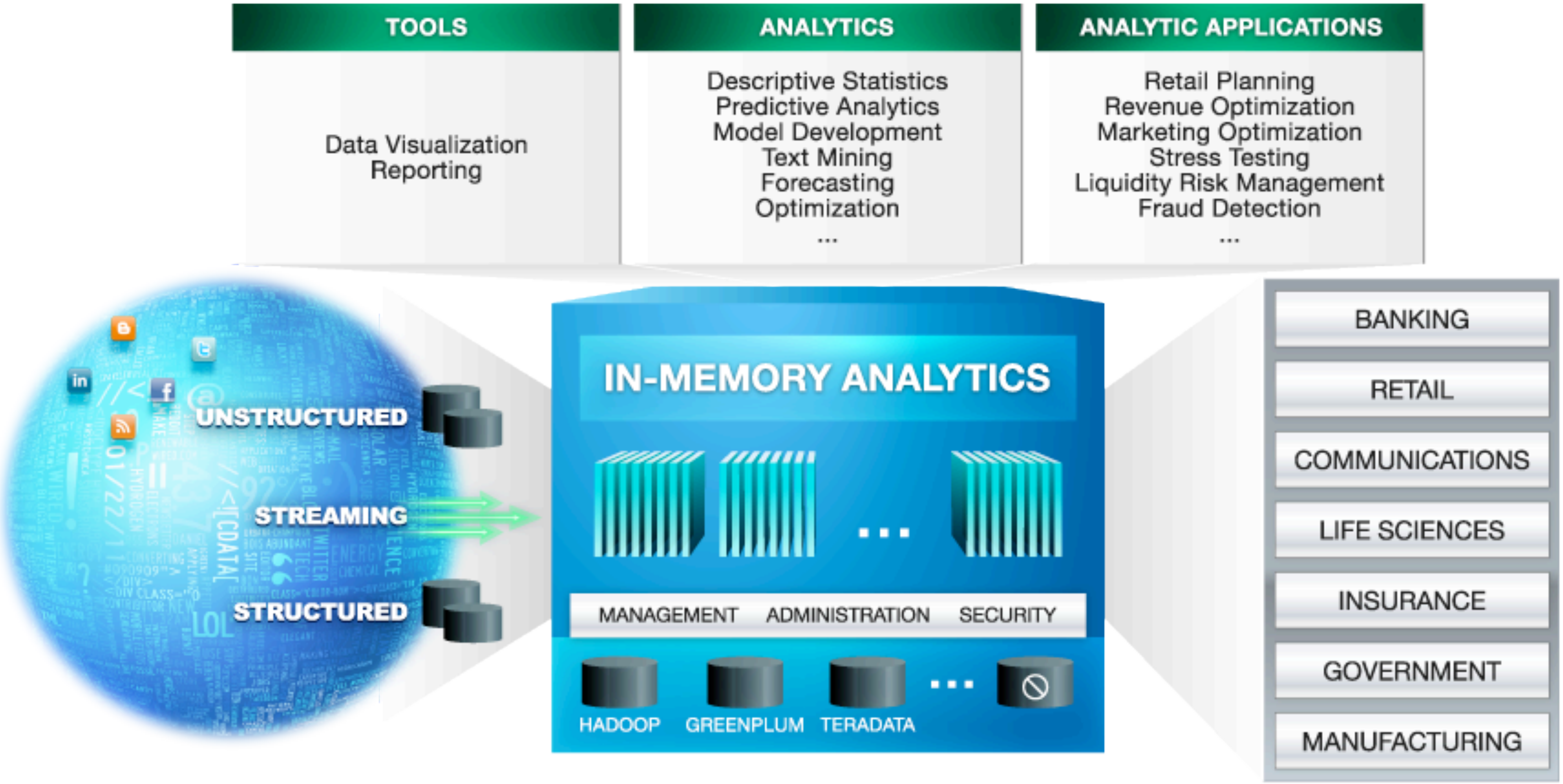
2. Analytics Query

3. Analytics Query Response

Framework Provider provides Processing and Data Store supporting interactive responses

Potential Standard: Analytic Query Language extending SQL

5. Example: In Memory Analytics



In Memory Analytics from SAS

Reference: http://www.sas.com/high-performance-analytics/how-does-it-work/in-memory.html

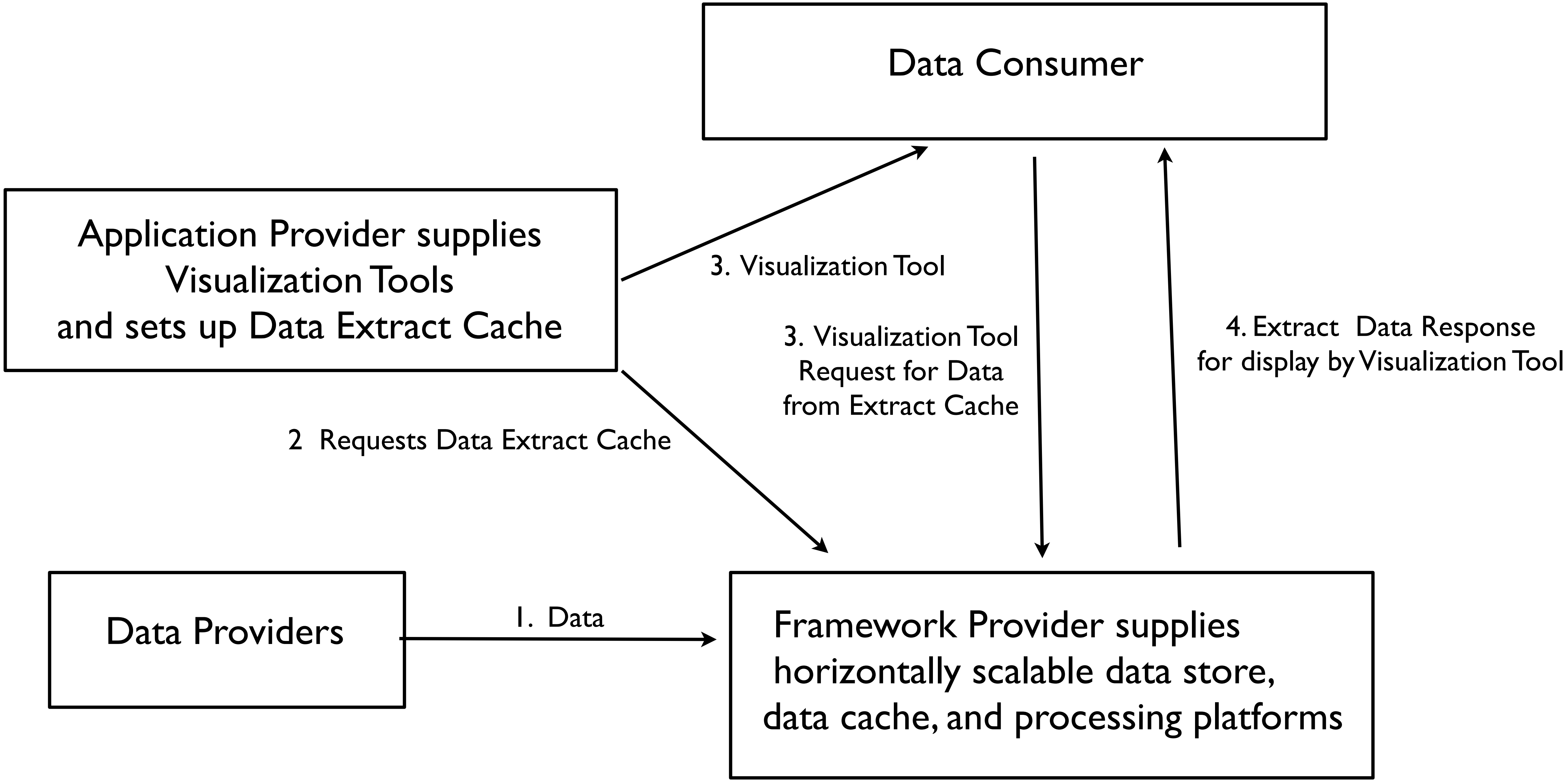
6. Generic Use Case

Use Case: Visualize data extracted from horizontally scalable Big Data stores

Mapping:

- Data provider supplies data.
- Application Provider supplies visualization tools (user interface, data cache, mappings from horizontally scalable data store to data cache)
- Framework Provider supplies horizontally scalable data store and processing platforms.
- Data Consumer uses the visualization tools to better understand the data.

6. Example: Visualization of Data Extracts



Data Consumer

Application Provider supplies Visualization Tools and sets up Data Extract Cache

3. Visualization Tool

3. Visualization Tool Request for Data from Extract Cache

4. Extract Data Response for display by Visualization Tool

2 Requests Data Extract Cache

Data Providers

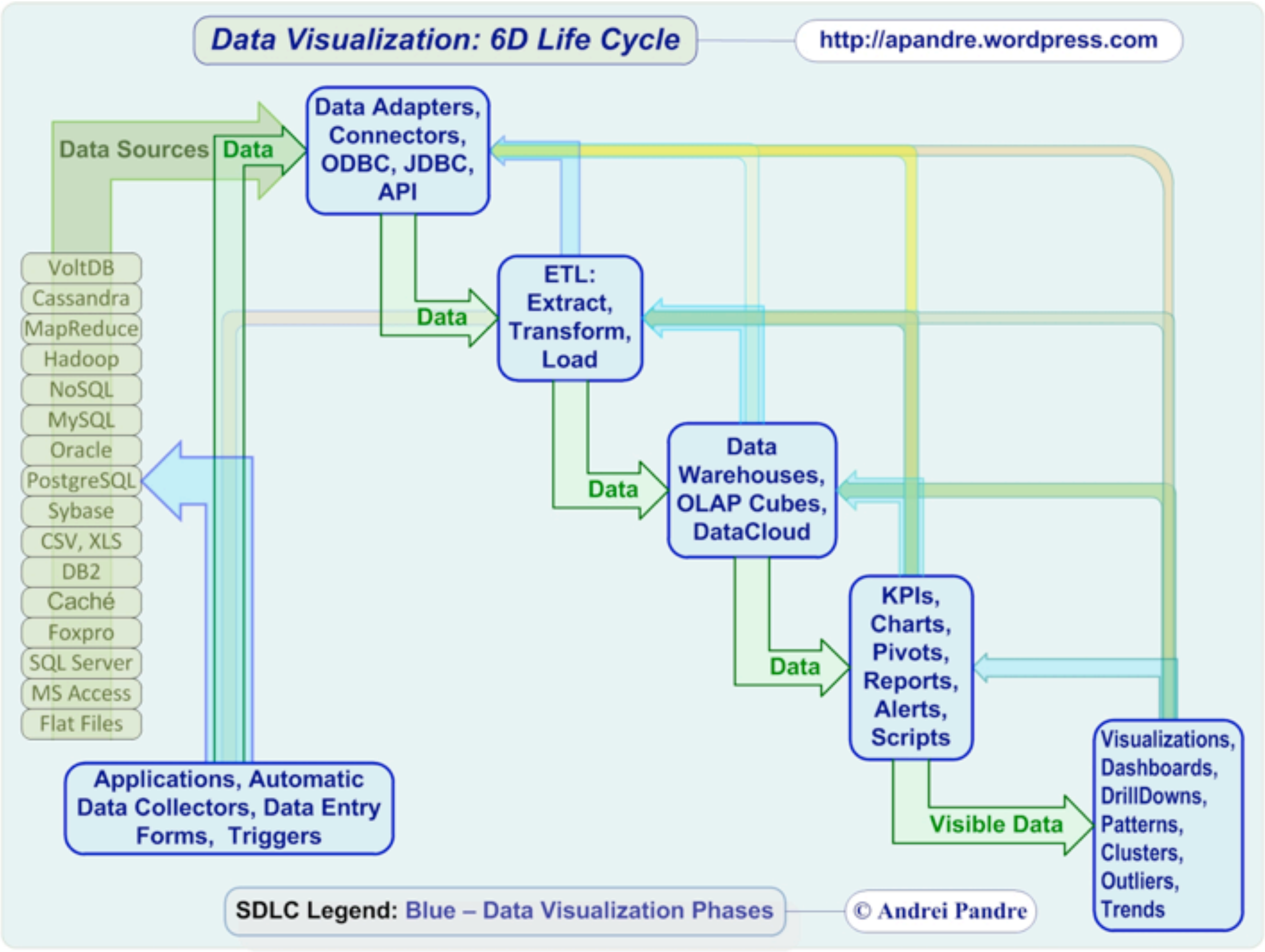
1. Data

Framework Provider supplies horizontally scalable data store, data cache, and processing platforms

Note that Data Provider sends the Data directly to the Framework Provider
This case can not be mapped directly to the Reference Architecture

Potential Standard: Visualization Request-Response Interface to Data Extract Cache

6. Example: Visualization of Data Extracts


Data Visualization: 6D Life Cycle
http://apandre.wordpress.com
Data Sources
Data
Data Adapters, Connectors, ODBC, JDBC, API
VoltDB
Cassandra
MapReduce
Hadoop
NoSQL
MySQL
Oracle
PostgreSQL
Sybase
CSV, XLS
DB2
Caché
Foxpro
SQL Server
MS Access
Flat Files
Data
ETL: Extract, Transform, Load
Data
Data Warehouses, OLAP Cubes, DataCloud
Data
KPIs, Charts, Pivots, Reports, Alerts, Scripts
Visible Data
Visualizations, Dashboards, DrillDowns, Patterns, Clusters, Outliers, Trends
Applications, Automatic Data Collectors, Data Entry Forms, Triggers
SDLC Legend: Blue – Data Visualization Phases
© Andrei Pandre

Data Visualization Process Flow

Reference: http://datavisualization.blog.com/page/5/

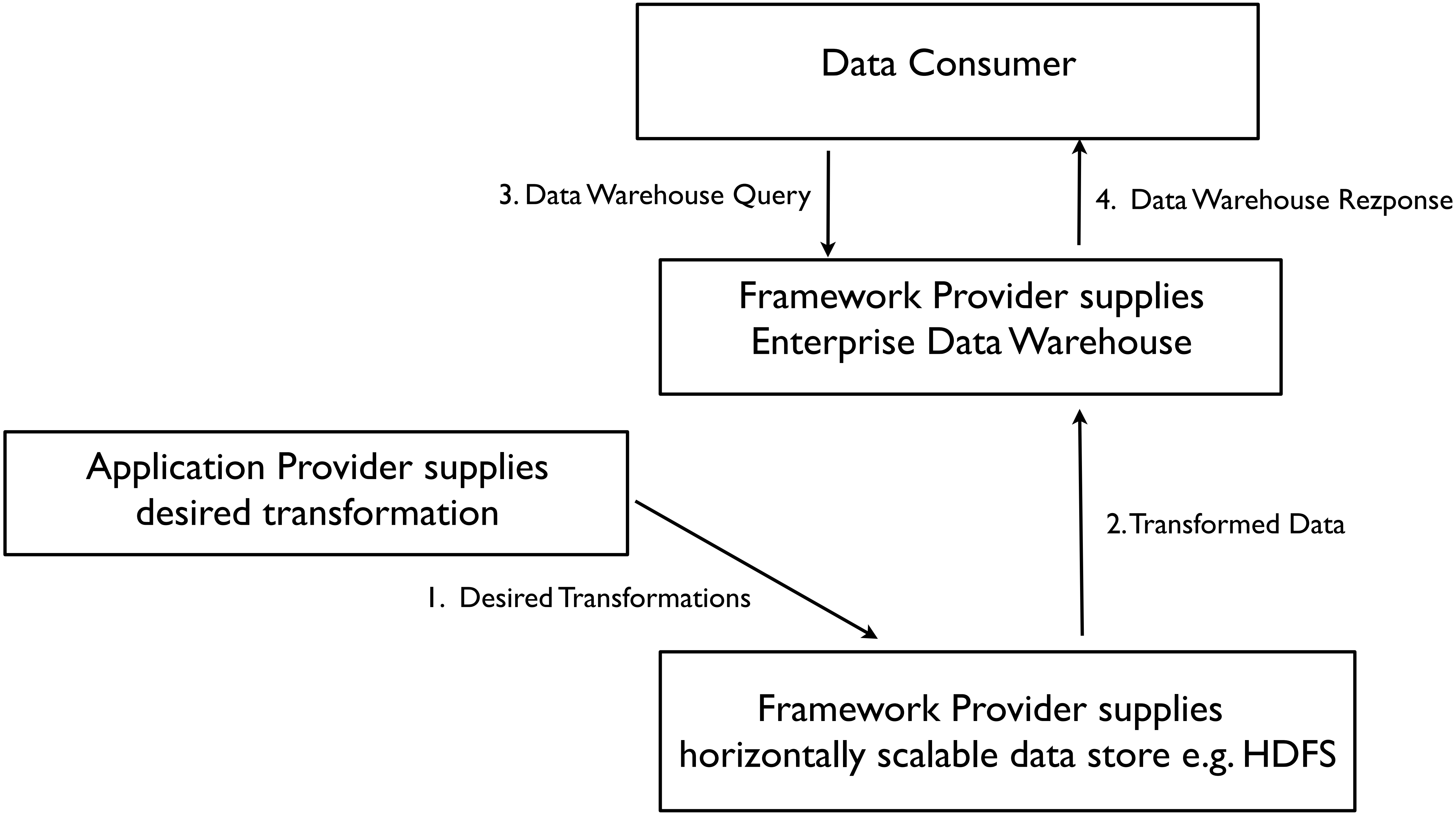
7. Generic Use Case

Use Case: Extract, process, and move data from a horizontally scalable data store into other target data stores (e.g Enterprise Data Warehouse or archival data store)

Mapping:

- Framework Providers supplies horizontally scalable data store and target data stores
- Application Provider supplies mapping to move data from horizontally scalable data store to target data stores
- Data Consumer uses target data stores

7. Example: Moving data from HDFS into EDW



Data Consumer

3. Data Warehouse Query

4. Data Warehouse Rezponse

Framework Provider supplies
Enterprise Data Warehouse

Application Provider supplies
desired transformation

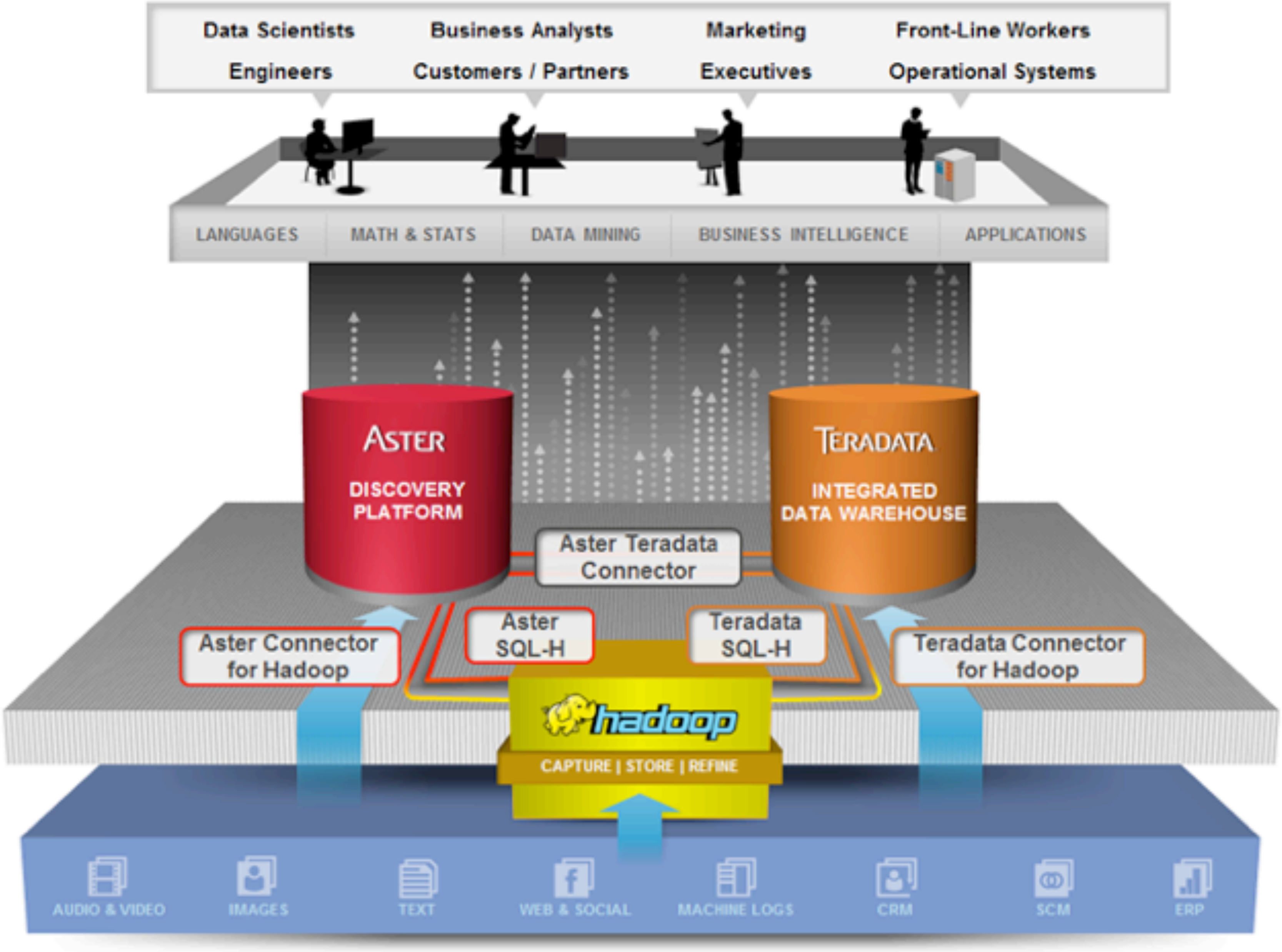
2. Transformed Data

1. Desired Transformations

Framework Provider supplies
horizontally scalable data store e.g. HDFS

Potential Standard: Declarative Description of Desired Transformations

7. Example: Moving data from HDFS into EDW



Moving data from HDFS to Teradata Data Warehouse and Aster Discovery Platform

Reference: http://www.asterdata.com/news/teradata-delivers-hadoop-data-to-the-enterprise.php

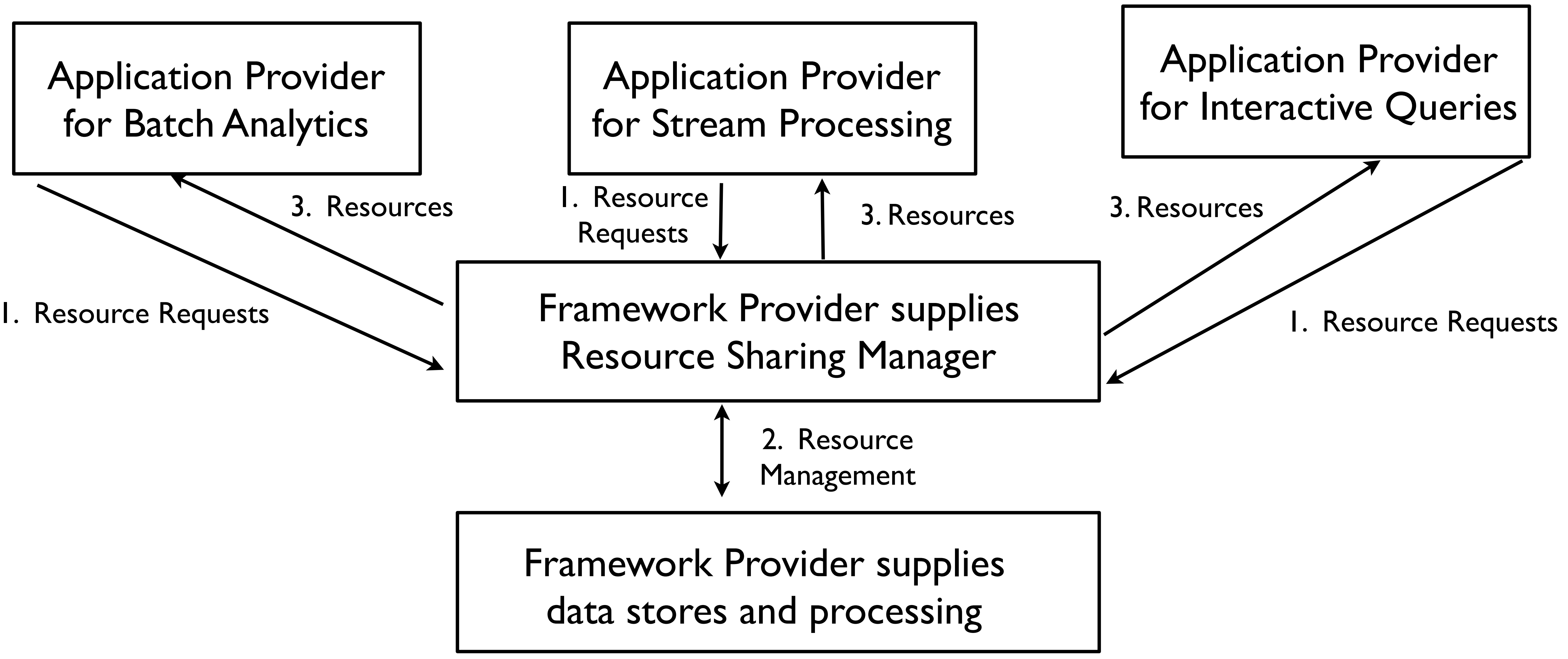
8. Generic Use Case

Use Case: Run multiple Big Data Processing (e.g. batch analytics, interactive queries, stream processing, network analysis) on top of shared horizontally scalable distributive processing and data store resources

Mapping:

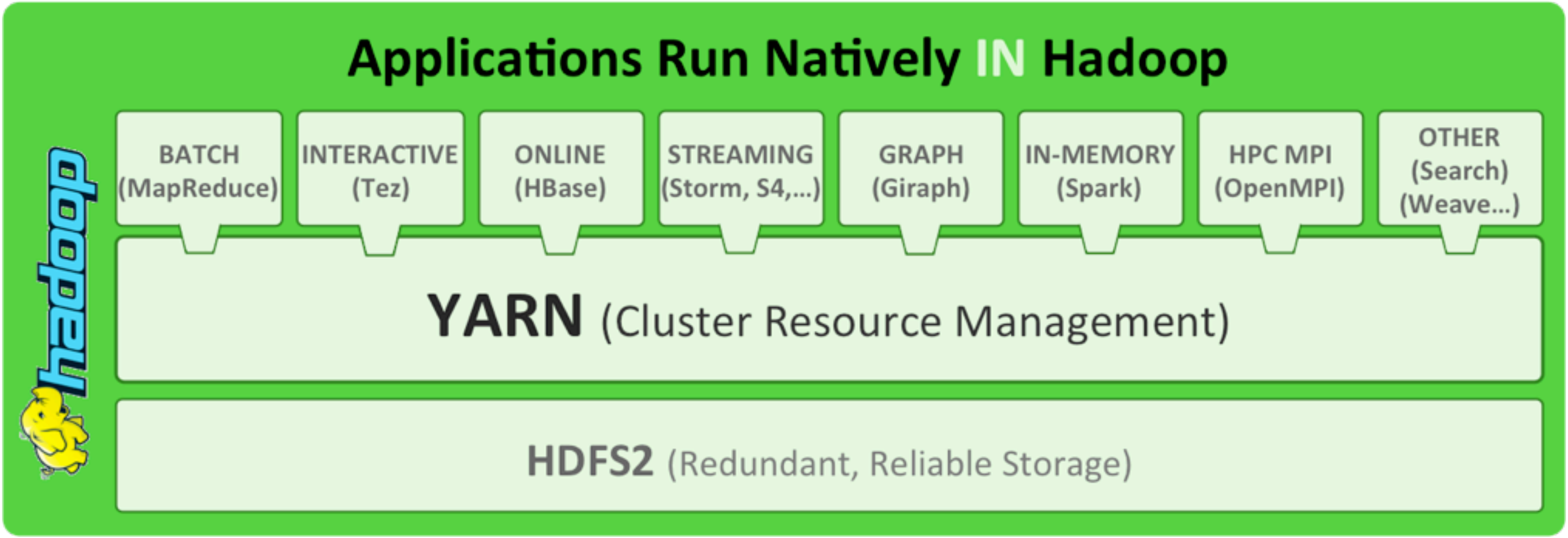
- Framework Providers provides shared processing and data resources and resource management frameworks
- Application Provider supplies applications to be run on shared resources

8. Example: Many Applications on Shared Resource


Application Provider for Batch Analytics
Application Provider for Stream Processing
Application Provider for Interactive Queries
3. Resources
1. Resource Requests
1. Resource Requests
3. Resources
3. Resources
1. Resource Requests
Framework Provider supplies Resource Sharing Manager
2. Resource Management
Framework Provider supplies data stores and processing

Potential Standard: Declarative Description for Resources and Resource Requests

Use Case 8. Resource Manager Example (YARN)



Applications Run Natively IN Hadoop

BATCH (MapReduce)

INTERACTIVE (Tez)

ONLINE (HBase)

STREAMING (Storm, S4,...)

GRAPH (Giraph)

IN-MEMORY (Spark)

HPC MPI (OpenMPI)

OTHER (Search) (Weave...)

YARN (Cluster Resource Management)

HDFS2 (Redundant, Reliable Storage)

Apache YARN Cluster Resource Manager

Reference: http://hortonworks.com/hadoop/yarn/

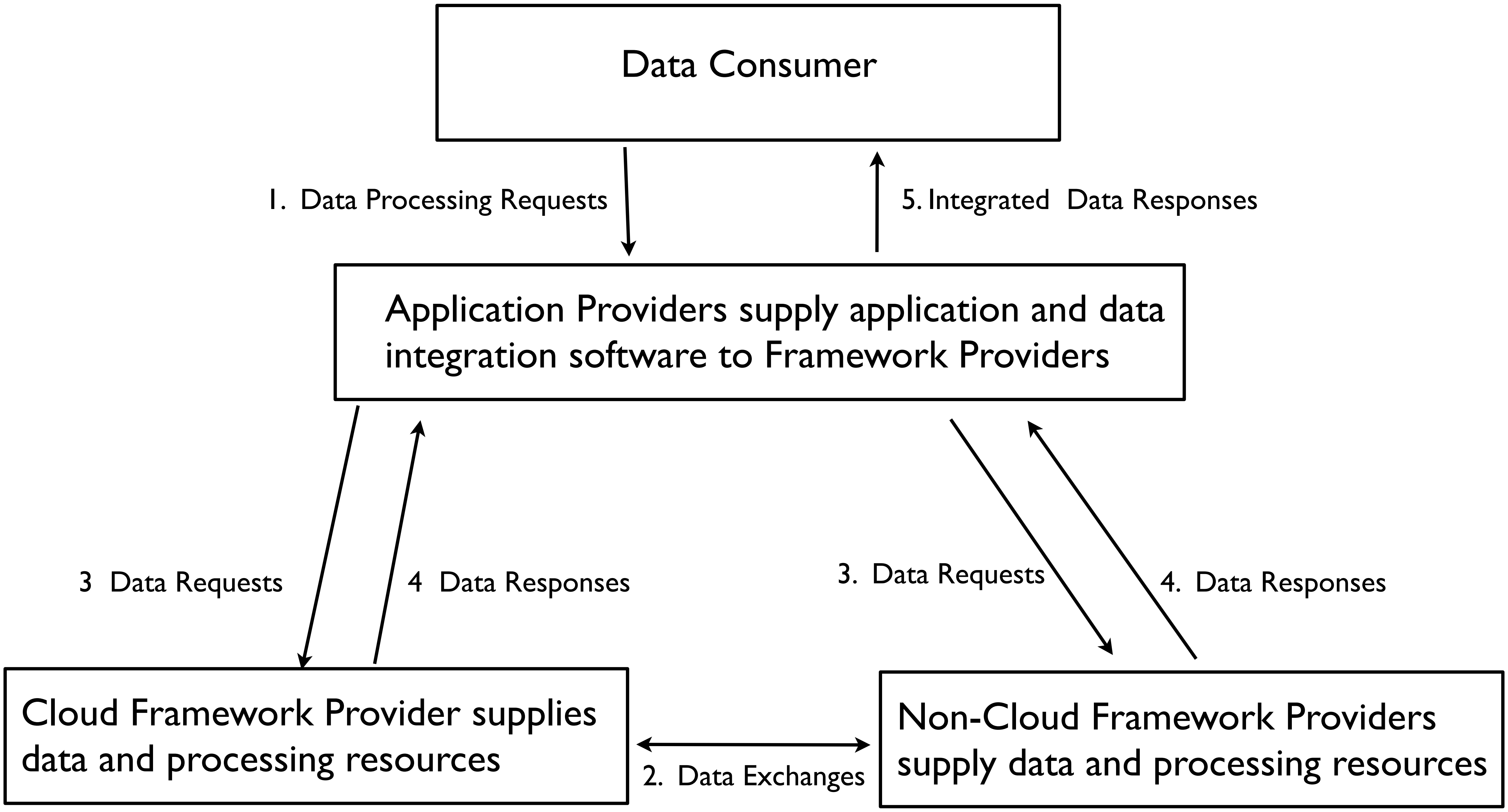
9. Generic Use Case

Use Case: Combine data from heterogeneous Cloud databases and non-Cloud data stores for analytics and data mining

Mapping:

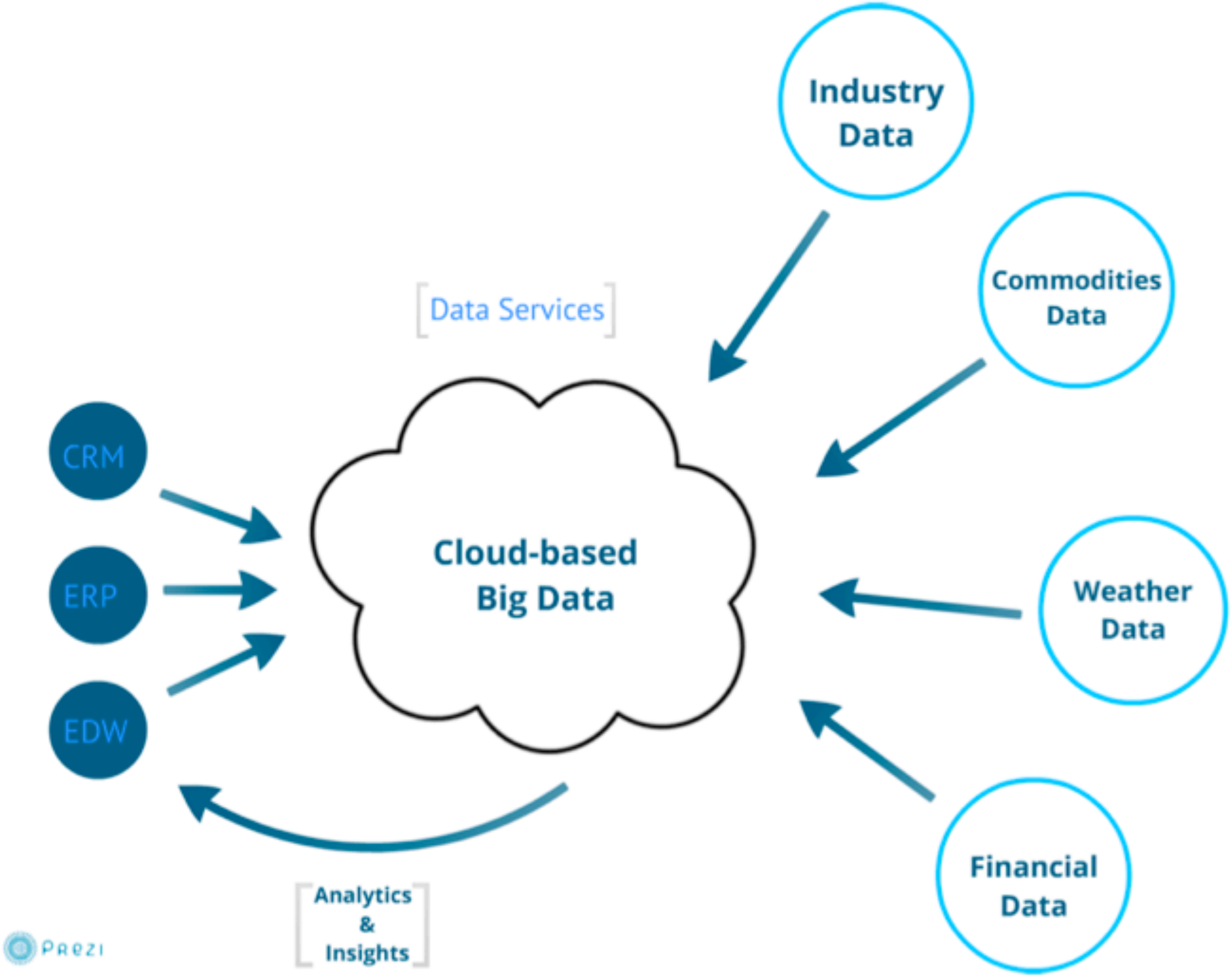
- Data Providers supplies data in data stores.
- Framework Provider hosts source data stores and archival data store.
- Application Provider supplies ETL, curation and packaging applications for archiving.

9. Example: Combining Cloud and Other Resources


Data Consumer
1. Data Processing Requests
5. Integrated Data Responses
Application Providers supply application and data integration software to Framework Providers
3 Data Requests
4 Data Responses
3. Data Requests
4. Data Responses
Cloud Framework Provider supplies data and processing resources
2. Data Exchanges
Non-Cloud Framework Providers supply data and processing resources

Potential Standard: Data Format and Metadata Standards to support Integration

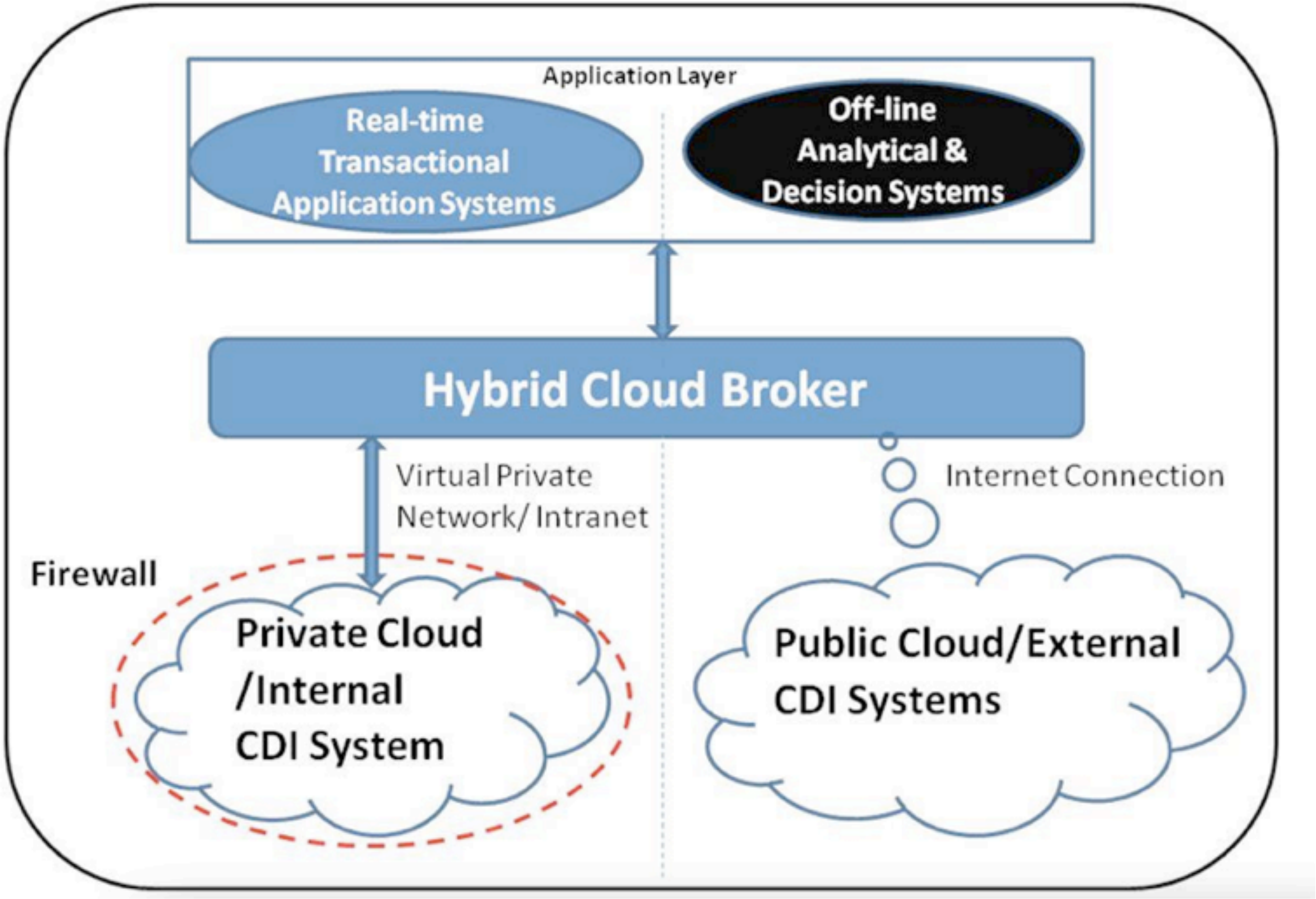
9. Example: Integrating Big Data into a Cloud



Integrating data from multiple sources into the Cloud

Reference: http://wikibon.org/w/images/2/20/Cloud-BigData.png

9. Example: Integrating Cloud and Internal Data


Application Layer
Real-time Transactional Application Systems
Off-line Analytical & Decision Systems
Hybrid Cloud Broker
Virtual Private Network/ Intranet
Internet Connection
Firewall
Private Cloud /Internal CDI System
Public Cloud/External CDI Systems

Note that Brokers can be supplied by Application Provider, Framework Provider, and/or System Orchestrator

Integrating data with a Hybrid Cloud Broker

Reference: searchbusinessintelligence.techtarget.in/tip/Benefit-from-customer-data-integration-CDI-on-cloud-platform

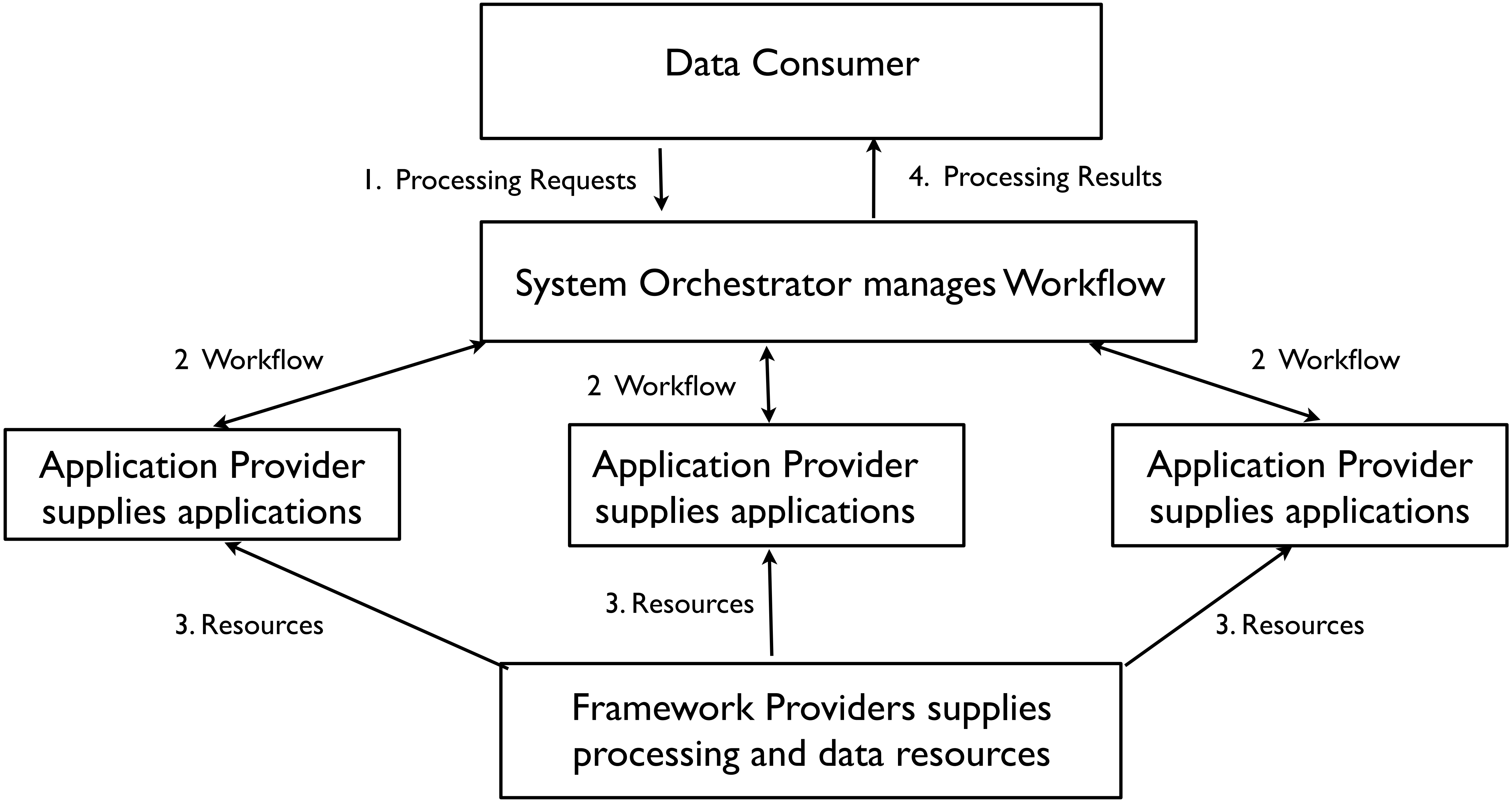
10. Generic Use Case

Use Case: Orchestrate multiple sequential and parallel data transformations and/or analytic processes using a workflow manager

Mapping:

- Data Consumer submits processing request to System Orchestrator
- System Orchestrator combines multiple use cases to provide a complete system for end-user from Data Provider through multiple data transformations and analytic processing
- Application Providers supply applications
- Framework Providers supply computing and data resources

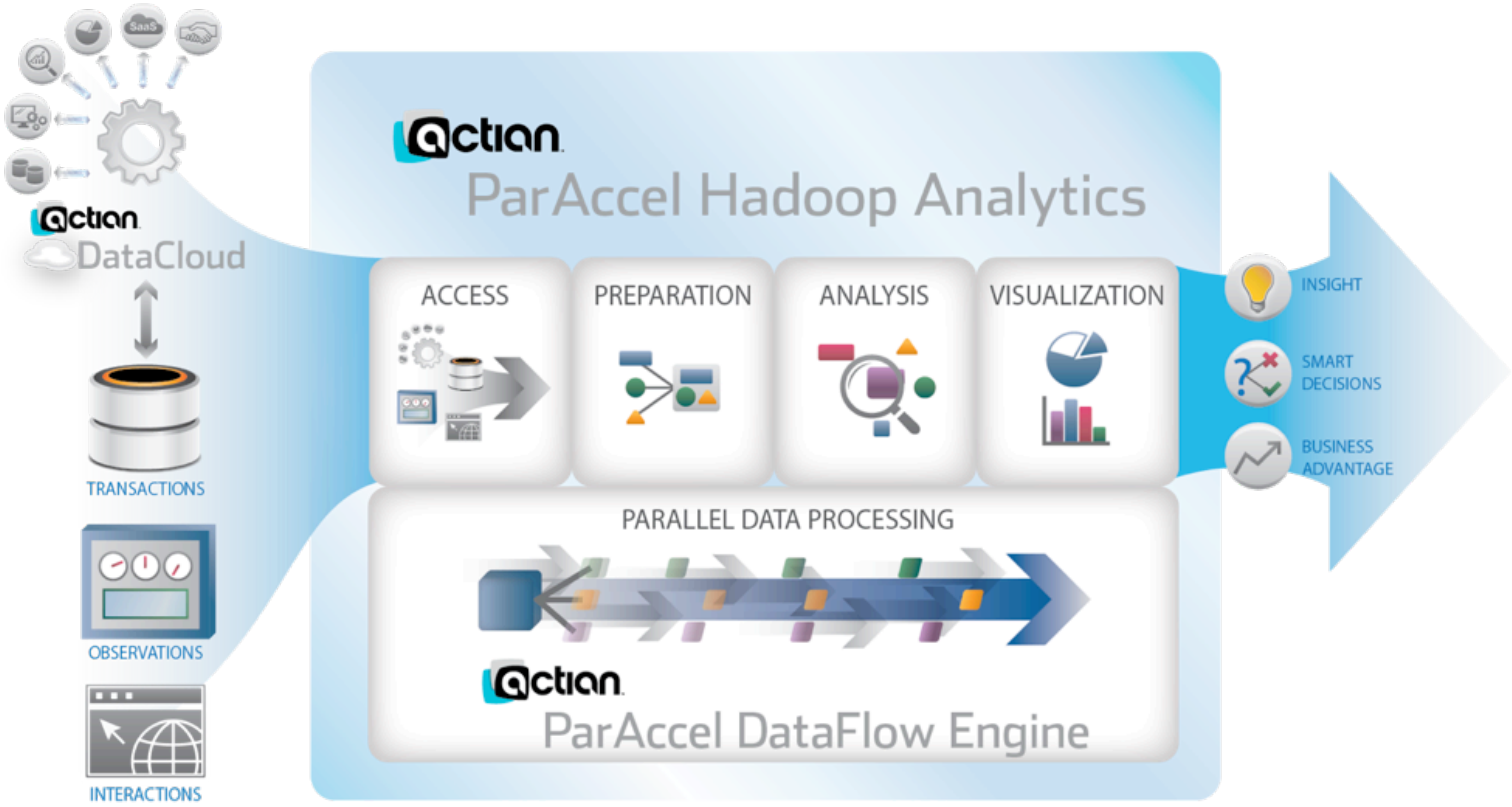
10. Example: System Orchestrator Runs Workflow


Data Consumer
1. Processing Requests
4. Processing Results
System Orchestrator manages Workflow
2 Workflow
2 Workflow
2 Workflow
Application Provider supplies applications
Application Provider supplies applications
Application Provider supplies applications
3. Resources
3. Resources
3. Resources
Framework Providers supplies processing and data resources

Note that Data Consumer interfaces directly to System Orchestrator which does not map to the Reference Architecture

Potential Standard: Declarative Descriptions of Processing Request and Workflows

10. Example: DataFlow Engine Runs Workflow



ParAccel Integrated Hadoop Analytics Workflow Process

Reference: http://bigdata.pervasive.com/Products/Big-Data-Analytics-RushAnalytics.aspx