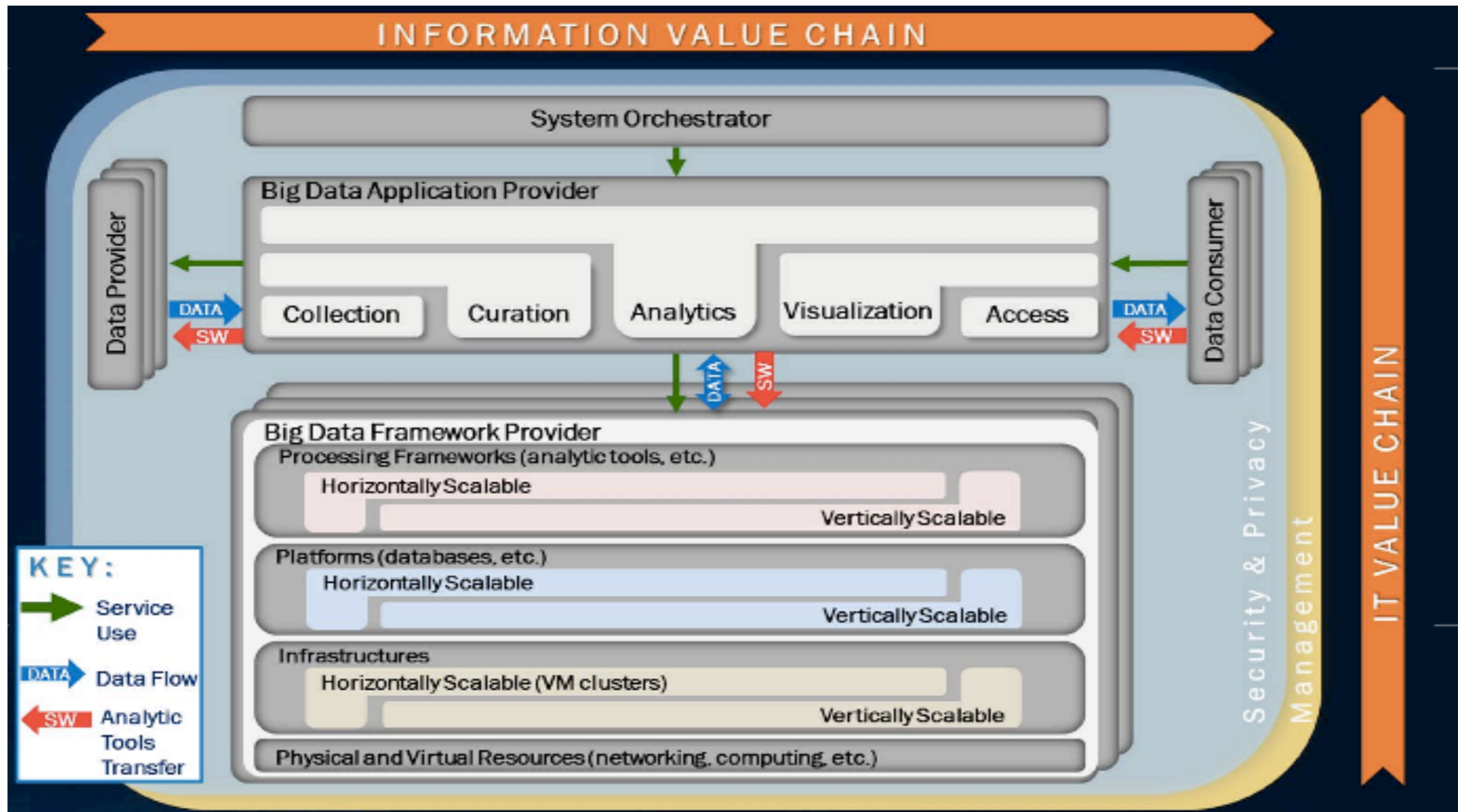


Generic Big Data Use Cases + Sample Mappings to Reference Architecture + Potential Standards

Bob Marcus
robert.marcus@et-strategies.com

NIST Reference Architecture



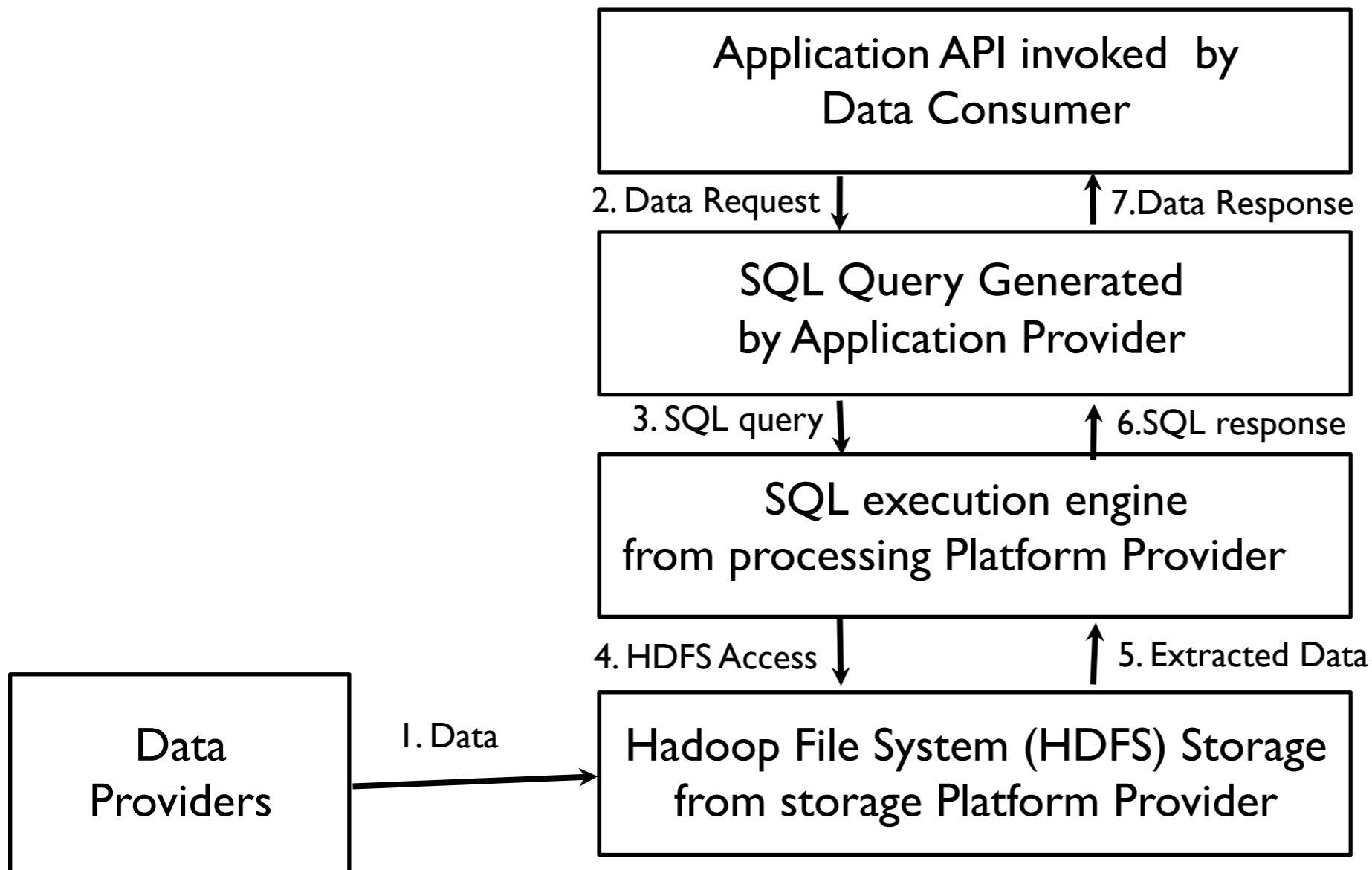
I. Generic Use Case

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.

I. Example: Interactive SQL Query to HDFS Data



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

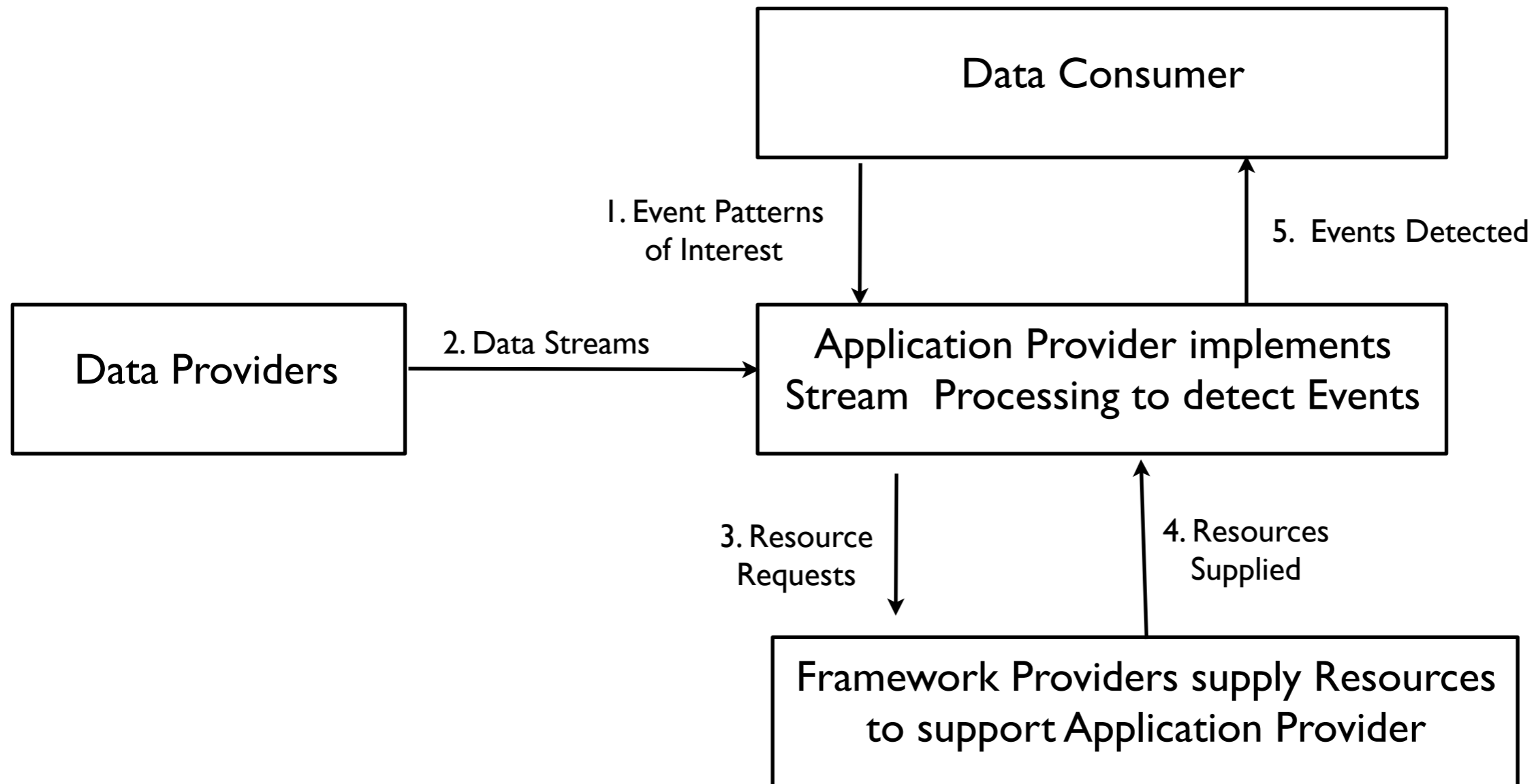
2. Generic Use Case

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



Potential Standard: Event Pattern Query Language and Event Description Language

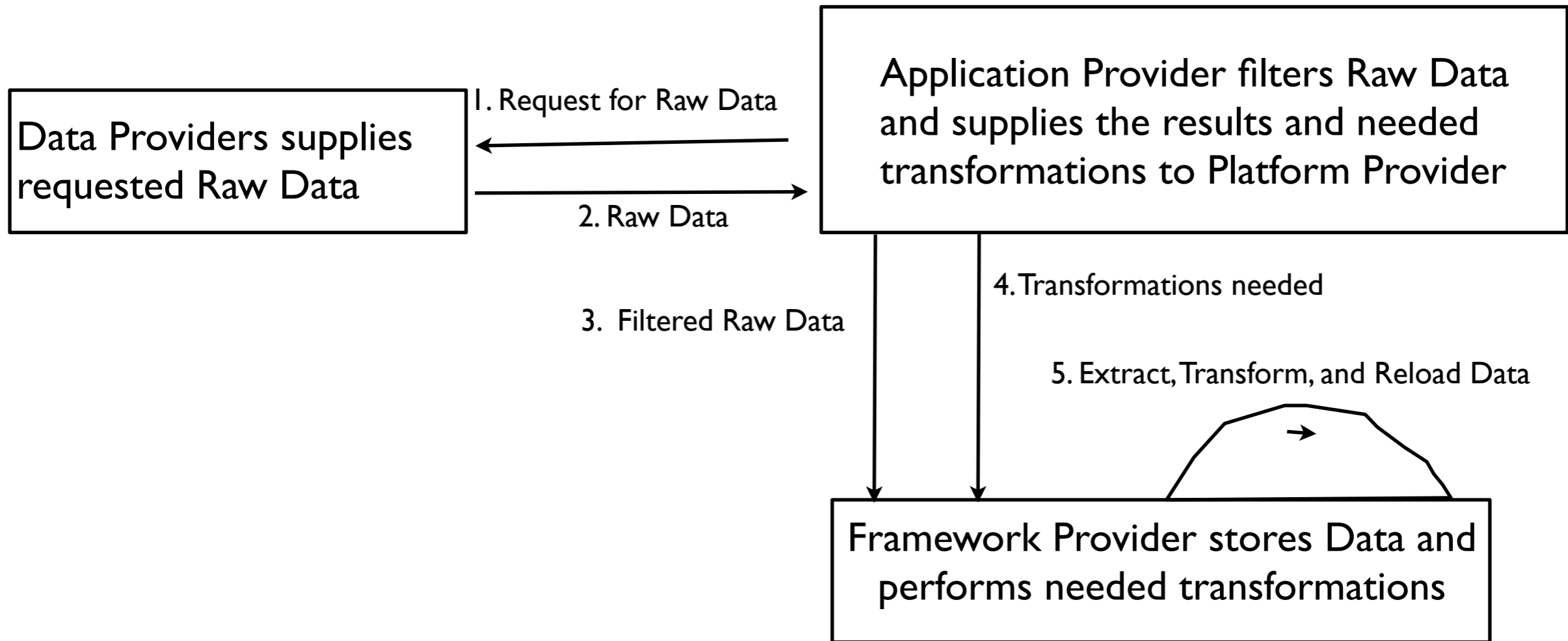
3. Generic Use Case

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



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 Descriptions Formats for Files, Data, and Metadata

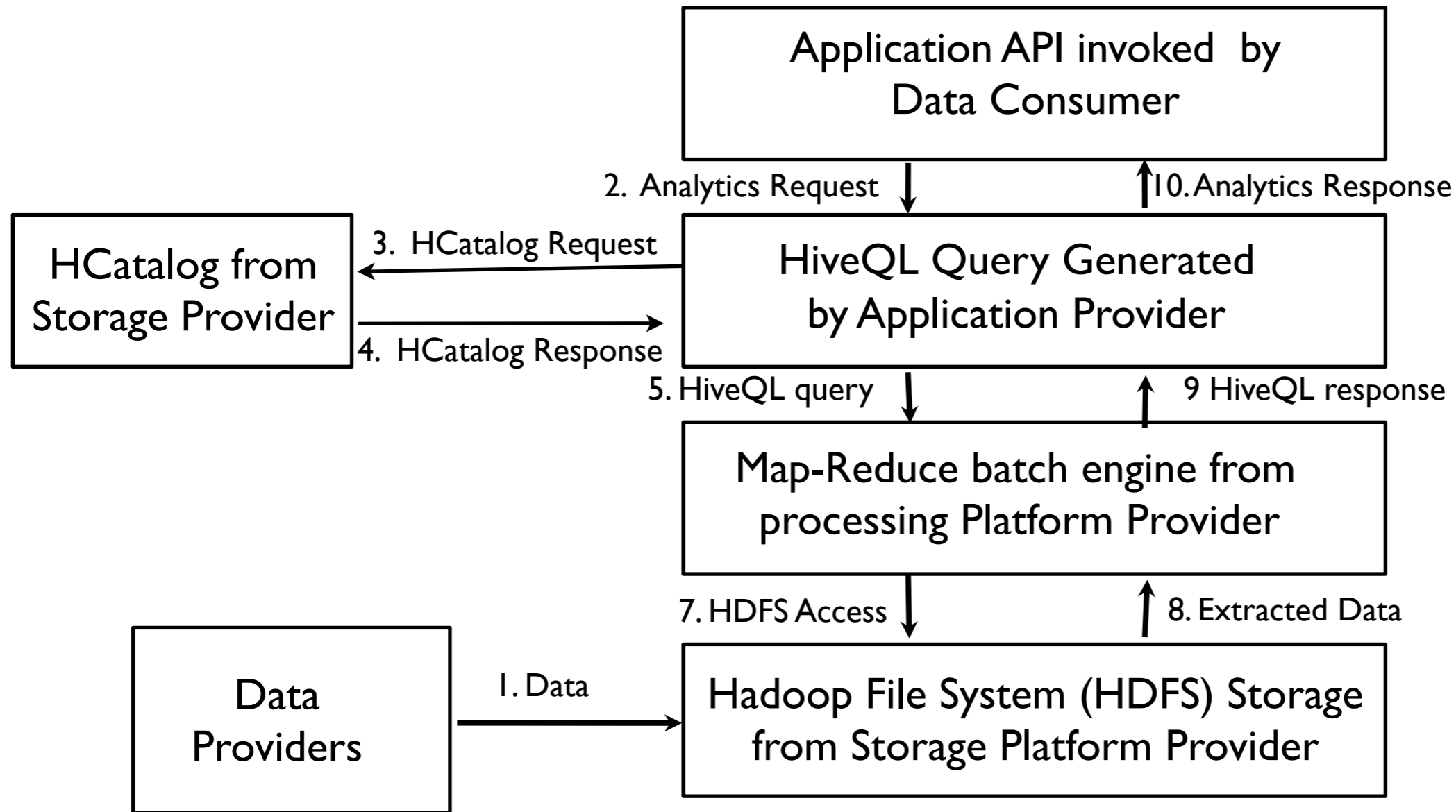
4. Generic Use Case

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



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)

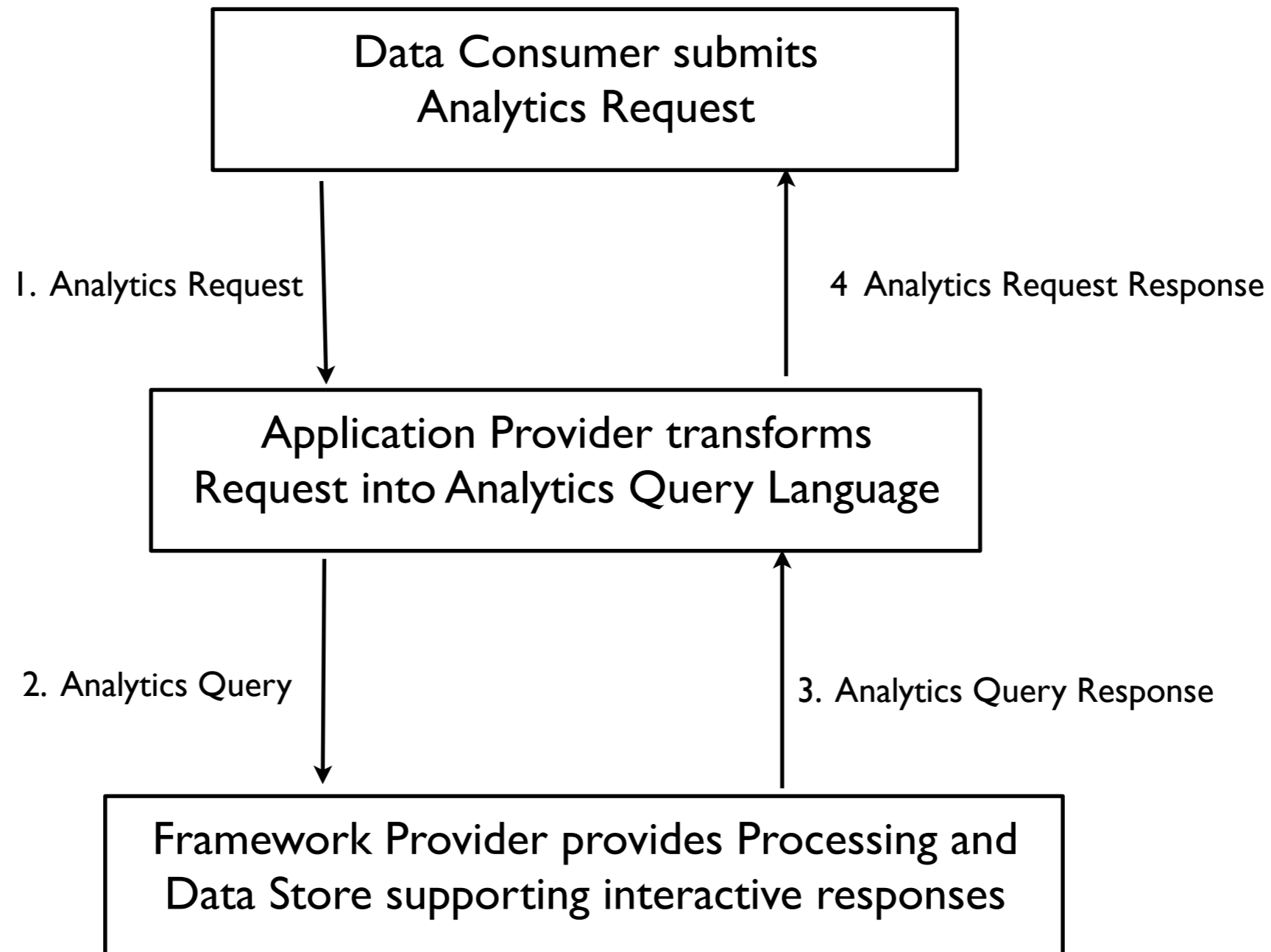
5. Generic Use Case

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



Potential Standard: Analytic Query Language extending SQL

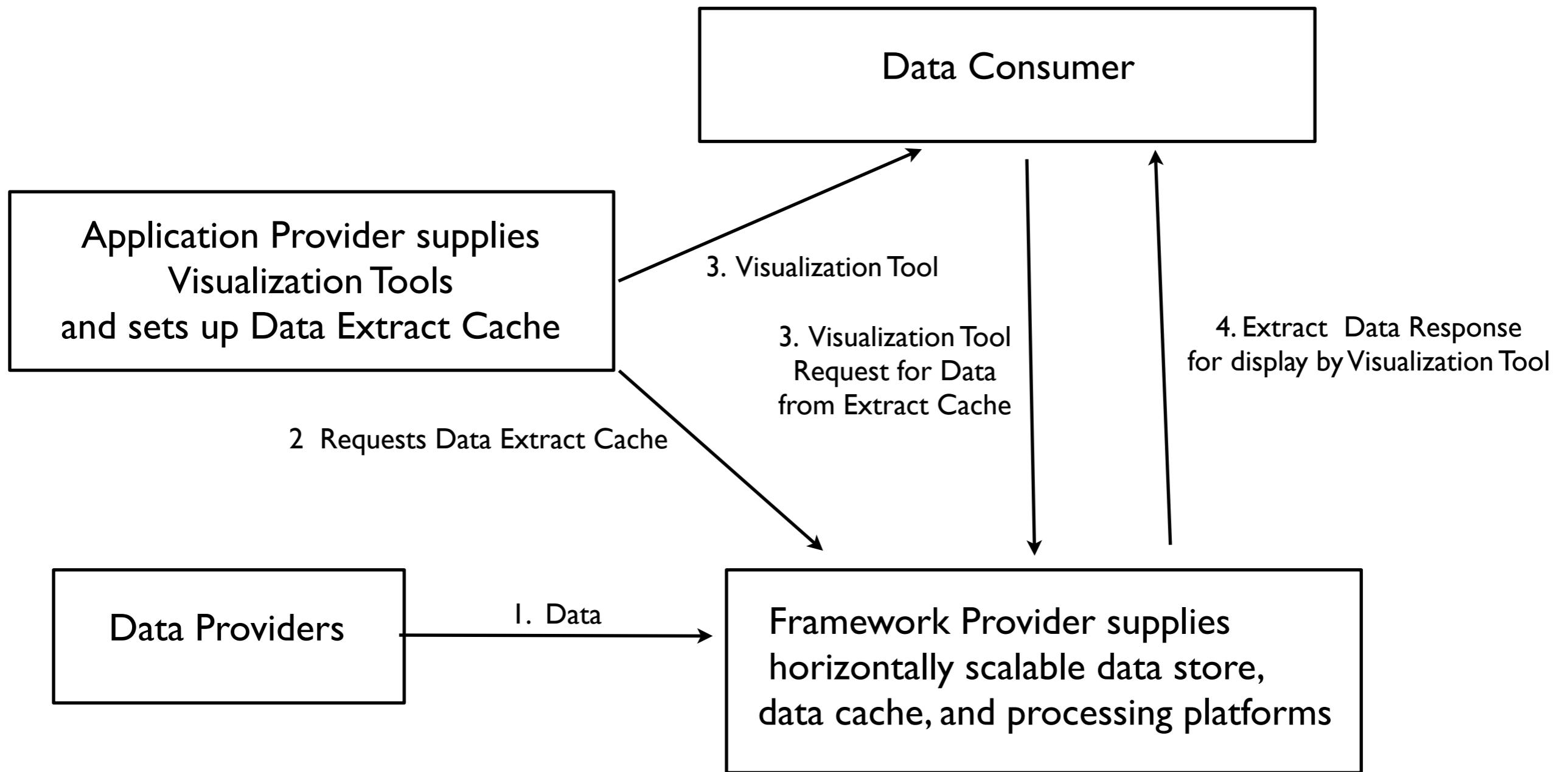
6. Generic Use Case

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



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

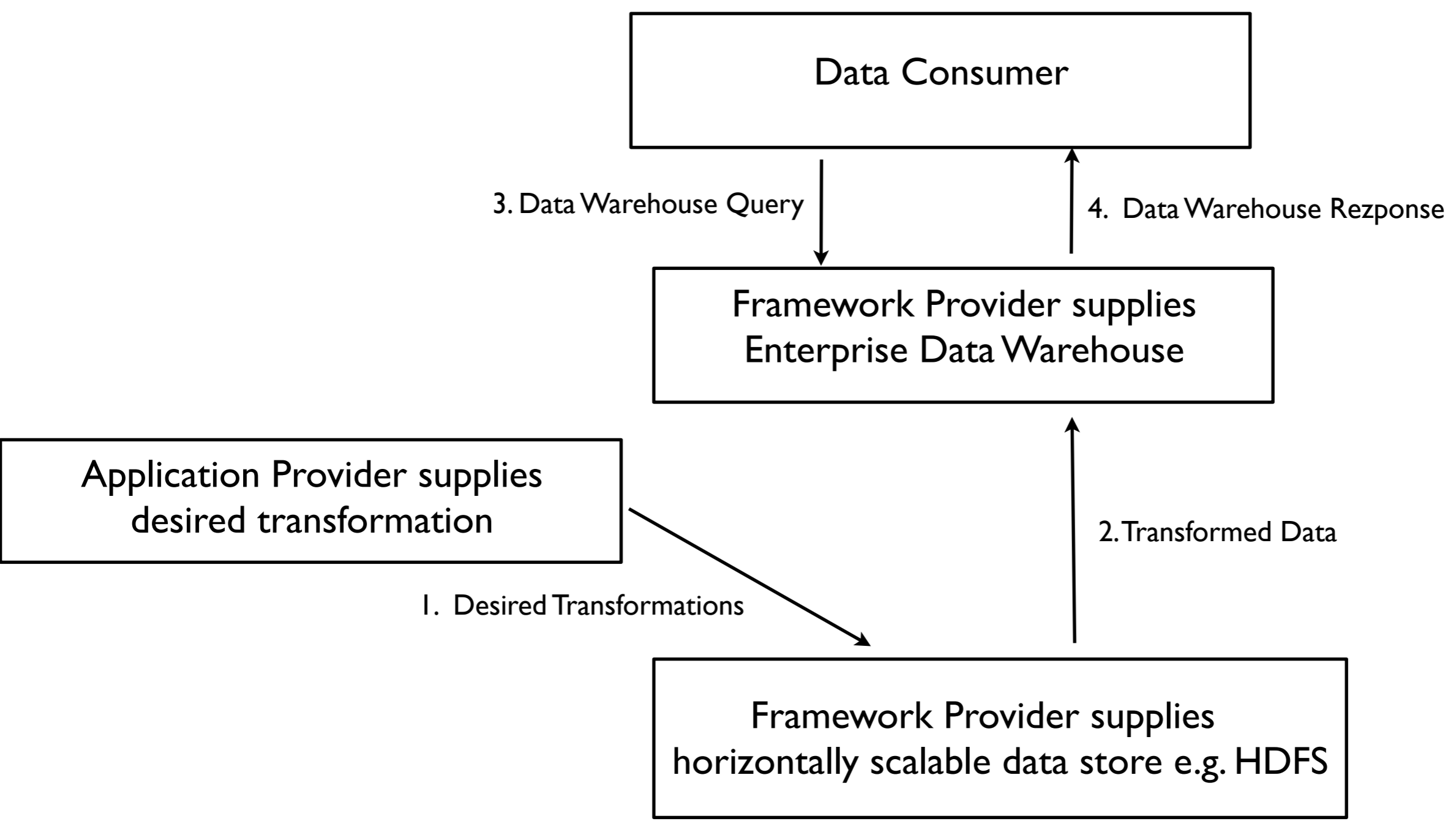
7. Generic Use Case

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



Potential Standard: Declarative Description of Desired Transformations

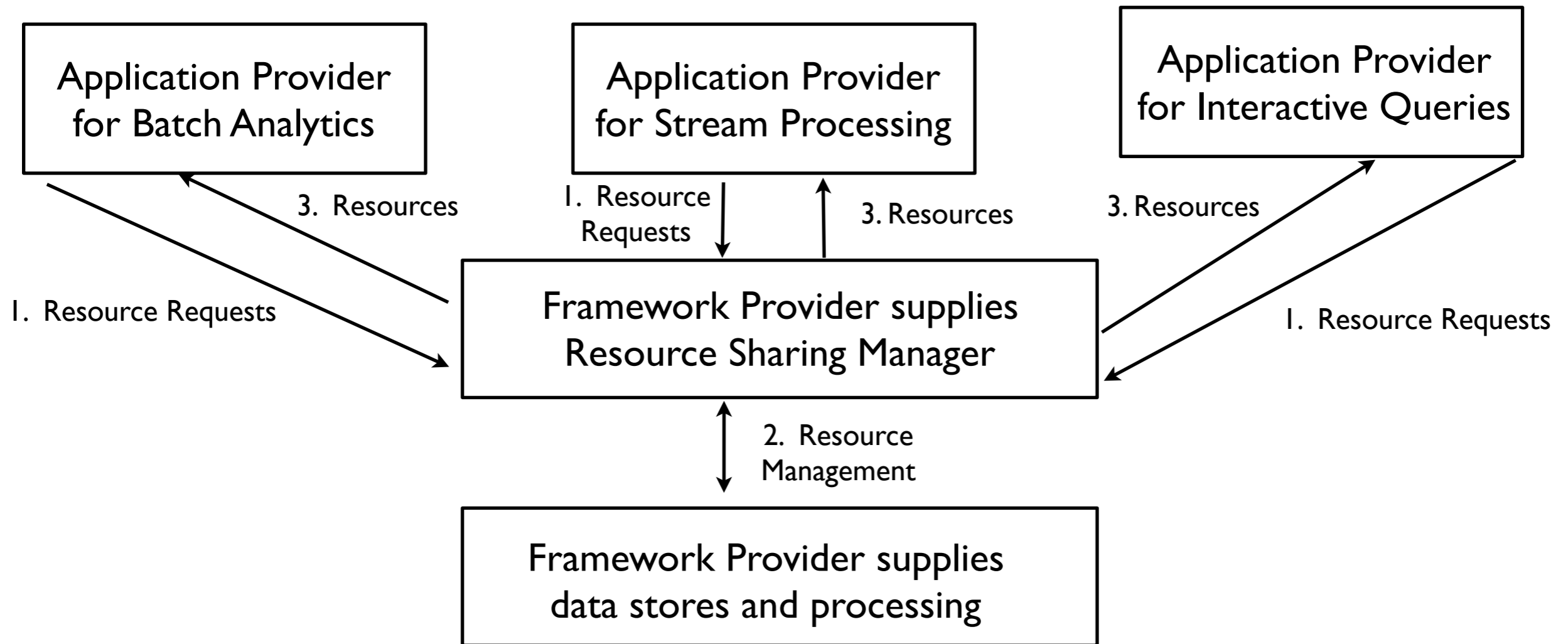
8. Generic Use Case

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



Potential Standard: Declarative Description for Resources and Resource Requests

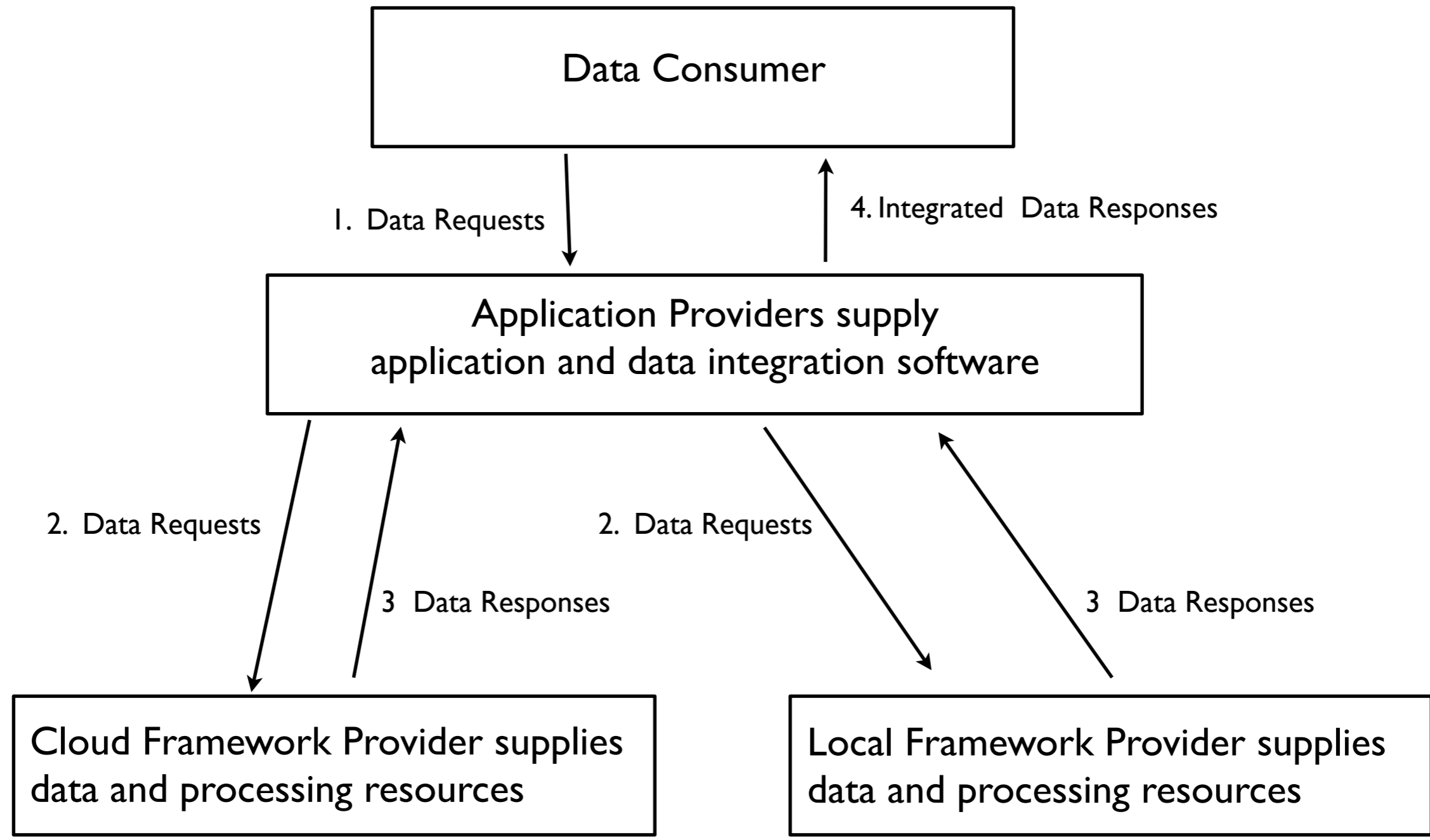
9. Generic Use Case

Use Case: Combine data from heterogeneous Cloud databases and on premise 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 Local Resources



Potential Standard: Data Format and Metadata Standards to support Integration

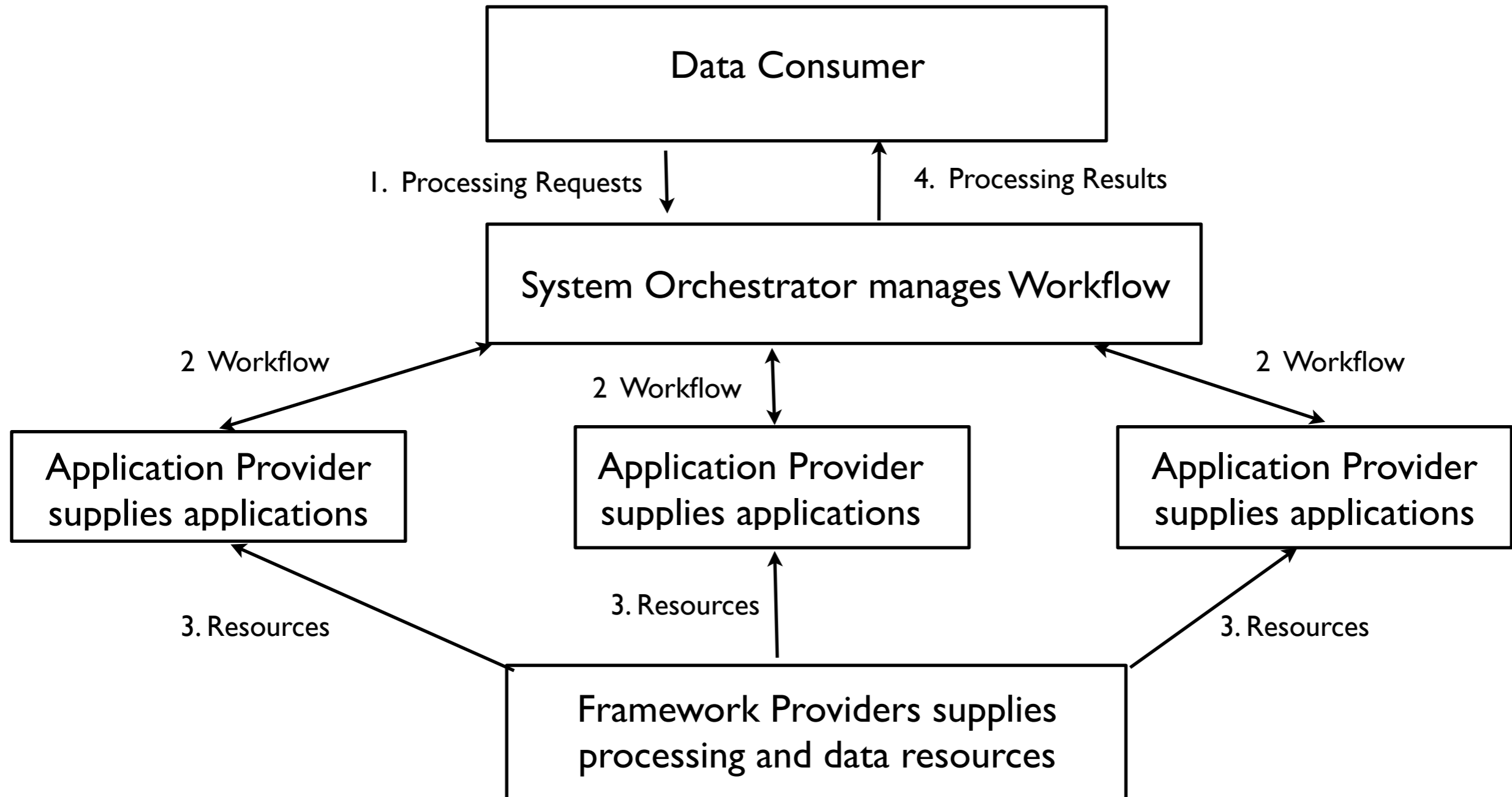
10. Generic Use Case

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



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