Data Pipelines & Provenance

Problem

As more applications become more API-centric and are assembled (and re-assembled), business context and output value take the form of data (small and large). Coordinating and following what is essentially a series of data value chains requires an over arching view that is not in a document, but handled such that it can be monitored, check pointed and even dynamically updated.

Solution

Publish APIs using API management (Application Blueprint*) that process information about data activities and which are then called by (or bundled into) application and service APIs to facilitate automated metadata management. Since, at the time of the event, the application has the information, auto update it as it happens. In addition, data movement and coordination requires over-arching orchestration that can be subscribed to. This could simply be a valet interface into the integration service. Next, leverage event processing and monitoring throughout the platform, including pre-processed service views summarizing activity in their respective domains of control (boundary, inspection, policy management, data services, data integration and API management). For security, data and application design patterns relationships can be auto determined and combined with the metadata. Patterns of data activity can be created as views with a dashboard or by observing data interactions. While it is encouraged that all data interactions use the data integration service (data pattern), even with extreme low latency, that level of introspection may be complete overkill for the task at hand. In those cases, APIs updated to metadata should suffice. At this point, this mega-pattern allows you to deliver an automated self-updating view of all data movement inside the environment and across clouds and ecosystems.

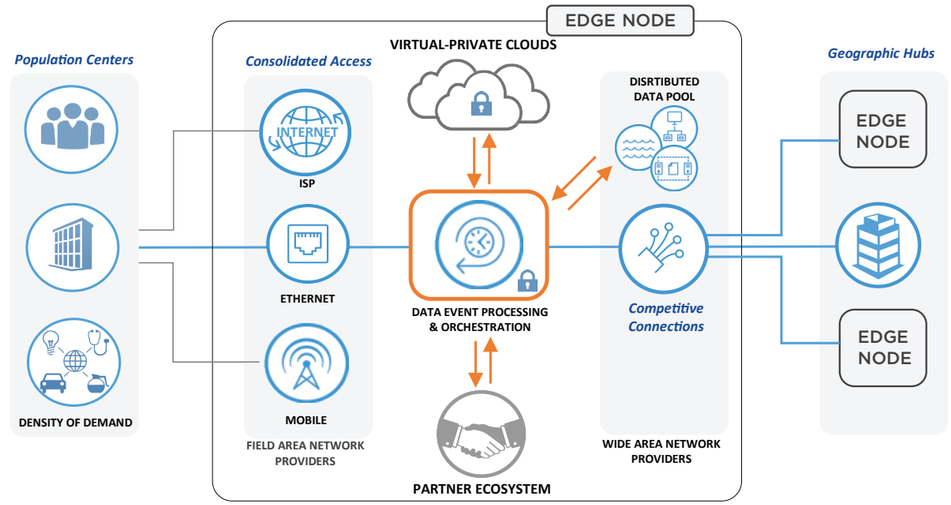
Constraints

1. As a whole, data flows and their associated metadata information are rarely documented or maintained.
2. While someone knows how the choreography of the application flow works, people change roles or leave and the knowledge goes with them.
3. Data movement, transfers, feeds, ETL, versioning, etc., all impact the health and performance of business operations but are not as visible or operationalized.
4. Therefore, consistency in data protection, security, quality, etc., may be lost as information traverses the environment.
5. Introduction of an ecosystem of partners and fast-paced changes to standup new business models quickly increases risk.

Steps

1. Apply data orchestration services to add scheduling and coordination of the other data services (listed below).
2. Leverage the data repository (self distributes globally) and local private storage (snapshots and access changes) for large data sets.
3. Leverage the data integration service for data services and data translation as needed.
4. Establish the provenance function to keep track of and publish APIs for automatic updates of metadata management and to query the service.
5. Leverage complex event processing (Application Blueprint*) to learn data relationships and trails, as well as construct views on data activities.
6. Update policy enforcement to flag anomalies of rogue data access and movement.

Reference View



Forces

- Controls must be balanced with performance requirements for speed of access and data transfer.
- Metadata information will become almost impossible to maintain (in rate of change and transparency and understanding).
- As the size and overall expense of data continues to rapidly grow, details on who is generating and using data will need to be well understood. Furthermore, its use needs to be correlated to business processes and value/benefits as well as inform company risk profiles.
- The proliferation of analytics, the demand for more data, and the ease at which data-driving functions are being cloned are forcing operational considerations to be solved first.

Results

- As an integrated family of optimized functions, the arduous challenge of master data management, risk and data security is made easier — without compromising availability, accessibility or performance.
- These services can be integrated with cloud services such that cloud use seamlessly satisfies operational requirements, updates dashboards and provides insights.
- Should a data breach occur — or more likely a colossal mistake — automatic checks and balances apply protection, with recovery as a fallback.
- Data expiration and HSM have not been covered, but are important. Much of what should be deleted based on policy or migrated to long-term archival is already there.

* Application Blueprint — IOAKB.com