Fybrik: a Policy Driven Open Framework for Managing the Use of Data

Speaker: Ronen Kat <ronenkat@il.ibm.com>

October 14, 2021
**Concepts**

**Fybrik: The challenges**

- **Data User**
  - It’s hard to get the data I need

- **Data Governance Officer/Steward**
  - I can provide guidance but cannot ensure governance policies are enforced

- **Operator / Admin**
  - Too many manual processes
  - I can’t optimize resource usage
**Fybrik: Moving non-functions capabilities to the platform**

**Business logic vs. Non-Functional**

**Business logic:**
The computation/algorithm which is the main objective of the data user

**Non-functional capabilities:**
Enforcing governance policies, technical connection to data stores, automatic movement of data, automation of data lineage collection, applying audit and data quality, optimizing access to data (e.g., caching)
<table>
<thead>
<tr>
<th>Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fybrik: Unlock value from data w/o scarifying security or speed</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Role</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data User</td>
<td>Fybrik helps abstract access to the data, enables the user to be agnostic to API, protocol and the location of the data or compute</td>
</tr>
<tr>
<td>Data Governance Officer/Steward</td>
<td>Fybrik enables enforcing governance over inputs and outputs in a consistent manner while enabling instrumentation for metrics, logs, audit, and data provenance independent of the workload</td>
</tr>
<tr>
<td>Operator / Admin</td>
<td>Fybrik enables to control and optimize the workload behavior by orchestrating data functions, data mobility, caching, copies and more</td>
</tr>
</tbody>
</table>
Fybrik: Declarative open approach

- **Concepts**
  - **Fybrik**: Declarative open approach
  - **APP Infrastructure focused**
  - **Workload focused**
  - **Data focused**
  - **POLICIES**
  - **Blueprint YAML**
  - **Compiles**
  - **Continuous deployment**
  - **Open library of modules**
    - Arrow/Flight
    - Masking and transformation
    - Data copy
    - Workload Isolation
    - Quality analysis
  - **Data User**
  - **Data Governance Officer/Steward**
  - **Operator/Admin**

- **Diagram**
  - **APP** → **Data**
  - **Data** → **YAML (Workload focused)**
  - **YAML (Data focused)** → **POLICIES** → **YAML (Infrastructure focused)** → **Compiles** → **Blueprint YAML** → **Continuous deployment** → **Open library of modules** → **Data**
Fybrik: Notebook demo

API version: app.fybrik.io/v1alpha1
kind: FybrikApplication
metadata:
  name: my-notebook
  labels:
    app: my-notebook
spec:
  selector:
    workloadSelector:
      matchLabels:
        app: my-notebook
appInfo:
  intent: fraud-detection
data:
  - dataSetID: "fybrik-notebook-sample/paysim-csv"
    requirements:
      interface:
        protocol: fybrik-arrow-flight
        dataformat: arrow

package data.api.authz

import data.data_policies as dp

transform[action] {
  description: "Redact sensitive columns in finance datasets"
  dpAccessType() == "READ"
  dp.dataset_has_tag("finance")
  column_names := dp.column_with_any_name({"nameOrig"})
  action = dp.build_redact_column_action(column_names[...], dp.build_policy_from_data)
Fybrik: Open architecture

- **Fybrik Core**
  - Credential Manager: e.g., Vault, k8s secrets
  - Data Catalog: e.g., Egeria, 3rd party
  - Data Governance Policy Manager: e.g., OPA

- **Fybrik Control Plane**
  - Credential Connector
  - Data Catalog Connector
  - Data Policy Connector

- **Fybrik Orchestration**
  - Blueprint Controller

- **Fybrik Data Plane**
Fybrik: Securing and controlling access to data
Fybrik Blueprint

Fybrik: Deploying data paths

And many, many more possibilities depending on the modules supported, governance decisions, etc....
Fybrik Eco-system

Fybrik: Building an eco-system around

Potential Integrations

3rd party modules
e.g., data quality, inspection, caching

trino  |  presto  |  Apache Drill  |  RAY  |  Spark  |  pandas

Amundsen  |  MARQUEZ  |  datashim

Flight  |  JDBC  |  S3  |  REST

HELP WANTED
Fybrick is a cloud native platform to unify data access, governance and orchestration, enabling businesses to securely manage enterprise data. By providing access and use of data only via the platform, Fybrick brings together performance and governance for data, greatly reducing the risk of data loss.

Contact: Ronen Kat <ronenkat@il.ibm.com>