Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 11, 2021

M. McBride
Futurewei
D. Kutscher
Emden University
E. Schooler
Intel
CJ. Bernardos
UC3M
D. Lopez
Telefonica I+D

July 10, 2020

Data Discovery Problem Statement
draft-mcbride-data-discovery-problem-statement-00

Abstract

If data is the new oil of the 21st century, then we need a standardized way of locating, capturing, classifying and transforming this raw data to generate insights and recommendations. Data, like oil, needs to be discovered and captured in order to be refined and valuable. While the topic of data discovery can be far reaching, this document focuses on the problem of actually locating data, throughout a network of data servers, in a standardized way.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

There are myriad proprietary and standardized ways of discovering networking devices and hosts. There are many solutions for discovering data within a database. There are proprietary, non-standardized, ways of discovering the data that may be stored throughout an environment of networking devices. We can discover information about the devices but can't locate and capture stored data in a standard way. With more networking devices storing collected data there needs to be a standard way of discovering the specific data needed amongst a potentially huge lake of databases.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Problem Scope

Data may be cached, copied and/or stored at multiple locations in the network on route to its final destination. With an increasing percentage of devices connecting to the Internet being mobile,

support for in-the-network caching and replication is critical for continuous data availability. There are data repositories throughout a modern network and there needs to be a standardized way to locating the repositories and discovering the desired data within.

There are many types of relational (SQL) and non-relational (NoSQL) data classification solutions. Existing database classification engines allow for scanning of a database. We are defining the problem, however, of having a standards based solution to discover first where the databases exist throughout a network and then where specific data objects are located.

Data discovery is likely to look different depending on if we are seeking global vs local discovery. Data discovery may be location-driven. A standard to find data may want to search for it in a more proximal fashion, i.e., find the data that matches the search that is nearest to a location.

There is so much data being created, processed, and migrated, that it may only sometimes get stored more permanently in a database. There is going to be slightly less permanent data that resides for a time in memory, so that it may be discovered and accessed quickly. It may be more dynamic and short lived. Although we refer to the data store as a database, it may reside entirely in memory, and/or it may be stored in some other non-SQL indexing technology.

Each database essentially provides a directory service for the data within them and that directory service can be viewed as metadata. There is the need to understand where the databases/data lakes/ pockets of data reside. The location of each data store is the first level discovery problem, and the details of the database's directory is the second level discovery problem.

Publish and subscribe approaches allow nodes to express their interest in specific pieces of data without knowing the location of the data. There might be sources of data to be discovered that might not produce the specific data desired by the subscribers (or not produce data with a specific format or frequency). The subscriber will want to find the publishers which send the desired data characteristics.

3. Existing Solutions

3.1. Proprietary

There are many existing proprietary database discovery solutions we can evaluate in order to understand what aspects we need to standardized. For instance there is IBM Cognos, Wipro Data Discovery

Platform (DDP), and Amazon Macie among many others. Macie, for instance, is a data security and data privacy service that uses machine learning and pattern matching to discover and protect data in AWS. The service allows you to define data types in order to discover and protect the data that may be unique to a use case.

3.2. Opensource

There are opensource data solutions such as from ScienceBase (https://sciencebase.usgs.gov/). The U.S. Geological Survey (USGS) is developing ScienceBase, an open source, collaborative, scientific data and information management platform. It provides current documentation about its structure, information model, services, directory and repository. sbtools uses an R (command line driven program used to find data within the platform) interface for ScienceBase.

Another solution is the Interplanetary File System (ipfs.io). IPFS is a distributed system for storing and accessing files, websites, applications, and data. IPFS is a peer-to-peer (p2p) storage network. Content is accessible through peers, located anywhere in the world, that might relay information, store it, or do both. IPFS knows how to find what you ask for via its content address, rather than its location. There are three fundamental principles to understanding IPFS:

- Unique identification via content addressing

- Content linking via directed acyclic graphs (DAGs)

- Content discovery via distributed hash tables (DHTs)

4. Use Cases

Here are some of the use cases which will benefit from standards based data discovery solutions:

- We need a standards based solution to discover the increasing amount of data being stored in various locations throughout a network including at the edge. We need a standard protocol set for doing this data discovery, on the device or infrastructure edge, in order to meet the requirements of many use cases. We will have terabytes of data on the edge and need a way to identify its existence and find the desired data. [I-D.mcbride-edge-data-discovery-overview] is focusing on this aspect of data discovery.

- We need a secure standards based solution for data discovery. Several of the proprietary secure data discovery solutions use machine learning and pattern matching to discover and protect the data. We need to incorporate existing, or new, ietf security solutions when discoverying data.

- We need a standards based solution for using named based solutions for data discovery. An Information Centric Networking (ICN) enabled network routes data by name (vs address), caches content natively in the network, and employs data-centric security. Data discovery may require that data be associated with a name or names, a series of descriptive attributes, and/or a unique identifier. NDN (Named Data Networking) can be applied to edge data discovery to make it much easier to extract data and meta-data by naming it. If data was named we would be able to discover the appropriate data simply by its name.

- We need a standards based way of discovering data in mobile wireless networks. Data could reside on the eNodeB or other wireless access infrastructure equipment in addition to residing on servers in the packet core.

5. IANA Considerations

N/A

6. Security Considerations

Data and metadata discovery are both a function of who asks for the data and in what context. The policies attached to the database and the metadata are going to dictate what view into the data that the system returns to the requester.

7. Acknowledgement

8. Normative References

[I-D.mcbride-edge-data-discovery-overview]
McBride, M., Kutscher, D., Schooler, E., and C. Bernardos, "Edge Data Discovery for COIN", draft-mcbride-edge-data-discovery-overview-03 (work in progress), January 2020.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

Authors' Addresses

Mike McBride
Futurewei

Email: michael.mcbride@futurewei.com

Dirk Kutscher
Emden University

Email: ietf@dkutscher.net

Eve Schooler
Intel

Email: eve.m.schooler@intel.com
URI: http://www.eveschooler.com

Carlos J. Bernardos
Universidad Carlos III de Madrid
Av. Universidad, 30
Leganes, Madrid 28911
Spain

Phone: +34 91624 6236
Email: cjbc@it.uc3m.es
URI: http://www.it.uc3m.es/cjbc/

Diego R. Lopez
Telefonica I+D
Don Ramon de la Cruz, 82
Madrid 28006
Spain

Email: diego.r.lopez@telefonica.com