Factsheet

SamKnows Shield – Big data

Big data

Big data
At SamKnows, as part of the SamKnows One platform we ingest and store huge amounts of data for processing and querying.

The measurement database is where all the test results are stored from all measurement agents globally. An extremely large volume of measurement data is generated daily, and this is only expected to increase. This volume of data warrants a different type of database to be used for the storage and querying.

Data collection
For our data collection infrastructure we have a variety of different services (depending on the agent type) which interact with measurement agents directly to collect their data and pass it through our ingestion pipeline into HDFS.

Data storage
We use Hadoop Distributed File System (HDFS) and Apache Hive for our primary metric data store allowing us to easily scale horizontally.

Querying
For querying we use a distributed SQL engine called Presto developed for analytical purposes by Facebook.

Infrastructure
Data storage & query processing workers
Our primary part of infrastructure which scales horizontally are 'Hadoop workers' which both stores data with an instance of a 'Hadoop DataNode', processes batch processing tasks on data with 'YARN NodeManager' and 'Hadoop MapReduce' services, and processing queries with an instance of a 'Presto Worker' service. Adding one new hadoop worker provides more data storage, improves performance of querying, increases querying capacity, and reduces querying contention.

A standard 'Hadoop worker' has the following specification, and they are added to existing clusters and remain within the same data centers as the rest of their clusters, often in very close proximity (nearby racks). This machine would provide ~11TB of usable storage capacity.

Component	Specification
CPU	Intel 2x Xeon E5-2630v3 - 16c/32t - 2.4GHz /3.2GHz
RAM	128GB DDR4 ECC 1866 MHz
Disk	3x4TB SATA

Analytics web servers
We have a number of web servers and general purpose workers per cluster operating in round-robin to serve analytics API requests and process jobs such as alerting checks. They also operate but with failover, in order to provide both high-availability and round-robin to provide load balancing. These are bare metal machines of the following specification.

Component	Specification
CPU	Intel Xeon E3-1270v6 - 4c/8t - 3.8GHz /4.2GHz
RAM	64GB DDR4 ECC 2400 MHz
Disk	2x450GB SSD NVMe

Data collection servers
Our data collection servers can handle approximately 150,000 measurement agents reporting data per server, however we recommend having n + 2 servers in order to

have high availability, including during maintenance windows. Our standard spec is as follows and can be either a VPS or bare metal machine:

Component	Specification
CPU	4vCPUs
RAM	8GB
Disk	400GB SATA
Bandwidth	10TB

Fault-tolerance

Each cluster has a large internal fault tolerance for servers/disks, generally aiming to be able to handle at least 10% of standard worker physical servers failing although often we can withstand even larger fault tolerance due to spread between rooms and racks and intelligent data replication.

All services within a cluster have minimum redundancy of n+1.

In addition to maintaining fault-tolerance within a cluster, SamKnows also maintains multiple clusters which is automated in the case of large scale failure. These clusters maintain provider-redundancy and geo-redundancy.

Scaling

Different parts of our infrastructure scale out depending on what is increasing.

- Data storage - requirements increase over time as more data is collected and will increase at a faster rate if more data is collected due to either increased number of measurement agents, more intensive test schedules or new functionality/improvements to the SamKnows One platform that involves collection of more environmental data)
- Data collection – requirements increase as more units/rows of data need to be collected per hour; the required storage capacity will increase as the number of measurement agents is increased, more intensive test schedules; wider geographic spread of devices or new functionality/improvements to the SamKnows One platform that involves collection of more environmental data.
- SamKnows One User Interface – requirements will be increased by a higher number of users utilising the platform, higher volume of API calls or users are more geographically spread
- Querying – requirements increase with increased query intensity (May be caused by new usage of the system by the ISP or new functionality developed by SamKnows), querying contention (How many people are using the SamKnows One analytics) or increased amounts of data collection (Increasing test schedule intensity, or increasing the amount of environmental data being collected).

Data Storage

The following table shows examples of how much additional storage space a single test might consume per day, excluding any environmental data; including storage in our big data store and associated redundancy and backups. This is based on the UDP Latency metric in particular and assumes test results always run (no cross-traffic prevention).

Component	Regularity of Test	Additional storage required per day
500,000	Once an hour	38GB
1,000,000	Once an hour	75GB
5,000,000	Once an hour	375GB
500,000	12 times per day	19GB
1,000,000	12 times per day	38GB
5,000,000	12 times per day	190GB

So therefore, assuming four tests of download, upload, and UDP latency; with download & upload producing 12 results per day, and latency producing 24 results per day, and no environmental data:

Agent Count	Storage per day	6 months of data	12 months of data
500,000	66GB	12TB	24TB
1,000,000	151GB	24TB	48TB
5,000,000	755GB	120TB	240TB

The SamKnows solution can scale to these kinds of measurement project sizes so long as time is given to provision extra hardware. On an ongoing basis we continue capacity planning at expected agent growth rates.