Factsheet

Android SDK

All software below is provided in binary form, unless noted otherwise.

What's included:

- Cross-traffic detection. The volume of cross-traffic produced by other apps on the device is captured along with the measurement results.
- Secure data collection and reporting. All measurement results are securely reported back to the SamKnows infrastructure.
- Test server discovery. The SamKnows backend will provide a list of candidate test servers, and the agent can determine the best server to use via a short latency check to each.
- Configurable tests. Users of the SDK can configure any parameter of the included test as desired.
- Documentation and source code for a sample application.
- Environmental data collection. The Android SDK captures a large amount of environmental data, including physical location, cellular network information, Wi-Fi information, and handset and operating system information.

Tests:

Speed tests

- Download (TCP)
- Upload (TCP)

Latency, loss and jitter

- Latency, packet loss and jitter

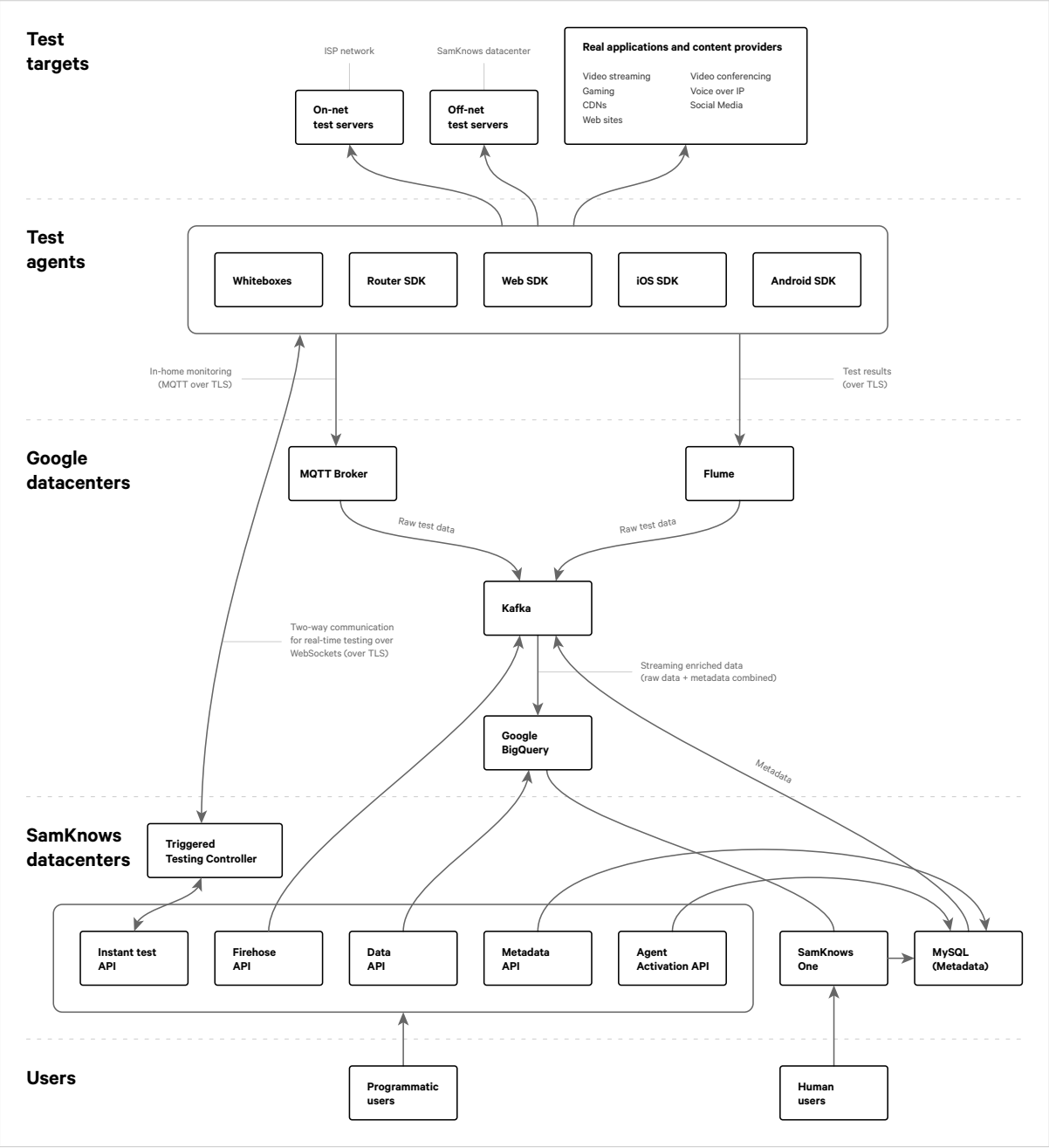
Video Streaming

- YouTube

Web browsing

- Web browsing

Overview of the components of the platform



Schematic of platform

Agents

The agents perform the measurements and carry out data collection. This includes both active measurements (against test servers and real applications) and passive environmental measurements. Agents may be provided in both SDK form (designed to be embedded inside a third-party product) and as a standalone product (e.g. the Whitebox).

Test servers

Test servers act as an endpoint for SamKnows agents to run measurements against. These test servers can be deployed “off-net”, which means outside of an ISP’s network, or “on-net”, which means inside an ISP’s network. No measurement data is stored on the test servers; they simply act as endpoints to generate and receive traffic.

Real applications and content providers

The agents also perform measurements against real applications and content providers. These include services for video streaming, gaming, CDNs, websites, social media, voice over IP and video conferencing services.

Flume

Flume is an open source Apache project for handling high-volume batch data collection. Flume acts as our gateway for incoming data. It receives data over HTTP (over TLS), validates and authenticates it, and then publishes it on a Kafka topic.

MQTT Broker

The MQTT Broker receives MQTT data from Whiteboxes and CPE. We use MQTT for delivering high-frequency structured realtime data, such as in-home environmental data.

Kafka

Kafka is an event-streaming data store. We use it to stream realtime measurement data into it, perform some transformations on it (such as splicing in metadata), and then publishing it to one or more consumers of the data. The primary consumer of data from Kafka is BigQuery.

BigQuery

BigQuery is the proprietary Google big data store that we use for long-term storage of measurement data and metadata. This provides very high scalability. Anything that accesses historical measurement data will query it from BigQuery; this includes SamKnows One Analytics and the Data API.

MySQL

MySQL is a relational database that we use to user data, Whitebox data, CPE data, app data and metadata. This is a far small database than the one hosted in BigQuery, but receives a far higher volume of reads and writes for transactional data.

SamKnows One

SamKnows One provides our ISP, government, and consumer users with a user interface to access the measurement results and manage their devices.

Agent Activation API

The Agent Activation API allows you to activate or deactivate a CPE for testing. When activating a CPE, you can optionally specify the test schedule it should be assigned to and a TTL (time to live) before it reverts to an inactive state. Activating a CPE will consume one of your CPE licenses.

Data API

The Data API provides read-only access to raw and aggregated measurement results, similar to that which you will find in SamKnows One Analytics. This API is intended for clients who wish to integrate our data into their internal platforms or backend systems.

Metadata API

The Metadata API allows you to attach supporting metadata to Whiteboxes or CPE. This includes items such as ISP, package, or service tier and timezone. You can also create entirely custom metadata fields that are specific to your network. This metadata can then be used in SamKnows One Analytics when analysing measurement results.

Firehose API

The Firehose API provides realtime streaming of measurement data directly from our backend. Consumers of this API will specify a key for the data they wish to subscribe to, and they will be pushed

events in realtime after that. This API is currently in development.

Instant Test API

The Instant Test API allows tests to be executed remotely in realtime on a SamKnows-enabled device (such as a Whitebox or CPE) and have the result returned synchronously. The time taken for each test will vary with the test requested (e.g. a 5 second speed test will result in a total response time of slightly more than 5 seconds).

Triggered Testing Controller

The Triggered Testing Controller acts as a gateway between Whiteboxes/CPE and the Instant Tests / RealSpeed functionality. Each Whitebox/CPE maintains a persistent connection of secure WebSockets to the controller, meaning that the controller needs to handle many millions of concurrent connections.

Hosting locations

SamKnows datacenters

MySQL is hosted in London and Canada, with realtime replication for resilience and high-availability. Personally identifiable data is stored in this database, such as name, email address, IP addresses and physical shipping addresses.

SamKnows One, the Triggered Testing Controller, Instant Tests API, Data API, Firehose API, Metadata API and Agent Activation API are hosted in London, Frankfurt, Singapore, Sydney, Toronto, New York and California.

All of the above components are deployed in an active-active fashion across multiple datacenters, with DNS load balancing used to steer traffic to the nearest location and manage failover.

Google datacenters

SamKnows uses GCP (Google Cloud Platform) to host its core data pipeline. This includes Flume, the MQTT proxy, Kafka and BigQuery. Google's cloud provides the resilience and failover automatically.

Any Google region can be used to host the Kafka and BigQuery components of the data pipeline, which is where data is stored. By default, the Europe-west region is used for all SamKnows services, with persistent data stored in the UK.