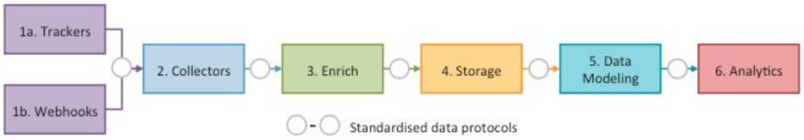
GitHub This repository Search Explore Features Enterprise Blog Sign up Sign in

snowplow / **snowplow** Watch 172 Star 1,915 Fork 485

Setting up Snowplow

Christophe Bogaert edited this page 19 days ago · 27 revisions

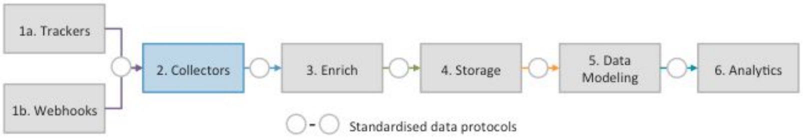
HOME > SNOWPLOW SETUP GUIDE



Setting up Snowplow is a six step process:

1. Setup a Snowplow Collector
2. Setup a Snowplow Tracker and/or Setup a Third-Party Webhook
3. Setup Enrich
4. Setup alternative data stores (e.g. Redshift, PostgreSQL)
5. Data modeling in Redshift
6. Analyze your data!

Step 1: Setup a Snowplow Collector



The Snowplow collector receives data from Snowplow trackers and logs that data to S3 for storage and further processing. Setting up a collector is the first step in the Snowplow setup process.

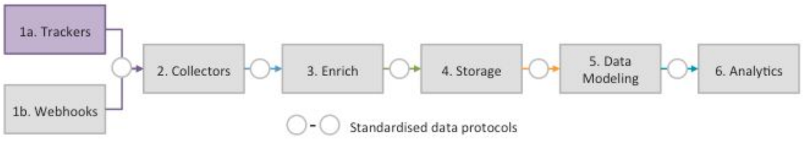
Setup a Snowplow collector now!

RootFile

Setup your collector? Then proceed to step 2: setup a tracker.

Step 2: Setup a Snowplow Tracker

Step 2a: Setup a Snowplow Tracker



Snowplow trackers generate event data and send that data to Snowplow collectors to be

Pages 201

HOME > SNOWPLOW SETUP GUIDE

Setup Snowplow

- Step 1: Setup a Collector
- Step 2a: Setup a Tracker
- Step 2b: Setup a Webhook
- Step 3: Setup Enrich
- Step 4: Setup alternative data stores
- Step 5: Data modeling
- Step 6: Analyze your data!

Useful resources

- Troubleshooting
- AWS sub-account setup
- IAM Setup
- Ruby and RVM setup
- Hosted assets

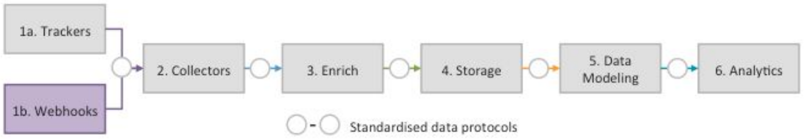
Clone this wiki locally

https://github.com/snowplow

captured. The most popular Snowplow tracker to-date is the JavaScript Tracker, which is integrated in websites (either directly or via a tag management solution) the same way that any web analytics tracker (e.g. Google Analytics or Omniture tags) is integrated.

Setup a tracker now! RootFile

Step 2b: Setup a Third-Party Webhook



Snowplow allows you to collect events via the webhooks of supported third-party software.

Webhooks allow this third-party software to send their own internal event streams to Snowplow collectors to be captured. Webhooks are sometimes referred to as "streaming APIs" or "HTTP response APIs".

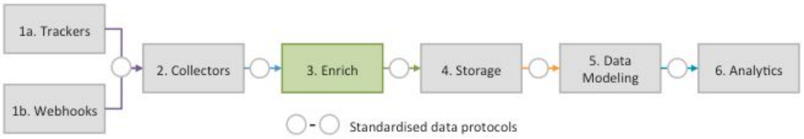
Setup a webhook now! RootFile

A note

Note: once you have setup a collector and tracker or webhook, you can pause and perform the remainder of the setup steps later. That is because your data is being successfully generated and logged. When you eventually proceed to step 3: Setup Enrich, you will be able to process all the data you have logged since setup.

Setup your tracker(s) and/or webhook(s)? Now proceed to step 3: setup Enrich.

Step 3: Setup Enrich



The Snowplow enrichment process processes raw events from a collector and

1. **Cleans up the data** into a format that is easier to parse / analyse
2. **Enriches the data** (e.g. infers the location of the visitor from his / her IP address and infers the search engine keywords from the query string)
3. **Stores the cleaned, enriched data**

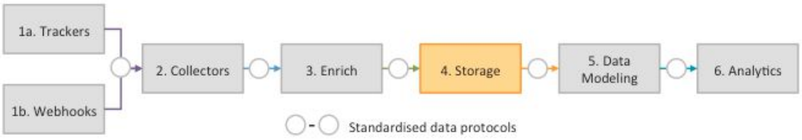
Once you have setup Enrich, the process for taking the raw data generated by the collector, cleaning and enriching it will be automated.

Setup Enrich now!

Setup Enrich? Proceed to step 4: setup the StorageLoader.

Step 4: Setup alternative data stores (e.g.

Redshift, PostgreSQL)



Most Snowplow users store their web event data in at least two places: S3 for processing in Hadoop (e.g. to enable machine learning via Mahout) and a database (e.g. Redshift or PostgreSQL) for more traditional OLAP analysis.

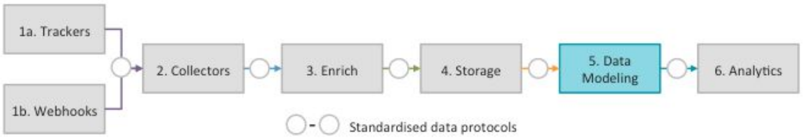
The StorageLoader is an application to regularly transfer data from S3 into other databases e.g. Redshift. If you **only** wish to process your data using Hadoop on EMR, you do not need to setup the StorageLoader. However, if you would find it convenient to have your data in another data store (e.g. Redshift) then you can set this up at this stage.

Setup alternative data stores!

RootFile

Setup the alternative data stores? Then proceed to step 5: data modeling.

Step 5: Data modeling in Redshift



Once your data is stored in S3 and Redshift, the basic setup is complete. You now have access to the event stream: a long list of packets of data, where each packet represents a single event. While it is possible to do analysis directly on this event stream, it is common to:

1. Join event-level data with other data sets (e.g. customer data)
2. Aggregate event-level data into smaller data sets (e.g. sessions)
3. Apply business logic (e.g. user segmentation)

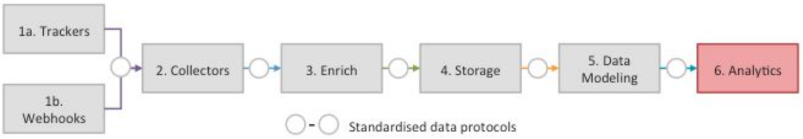
We call this process *data modeling*.

Get started with data modeling in Snowplow!

RootFile

Done with the data modeling? Then proceed to step 6: analyse your data.

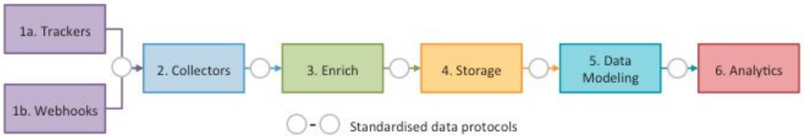
Step 6: Analyse your data!



Now that data is stored in S3 and potentially also Redshift, you are in a position to start analyzing the event stream or data from the derived tables in Redshift, if a data model has been built. As part of the setup guide we run through the steps necessary to perform some initial analysis and plugin a couple of analytics tools, to get you started.

Get started analysing Snowplow data!

The Snowplow setup is complete!



You now have all six Snowplow subsystems working!

Home | About | Project | Setup Guide | Technical Docs | Copyright © 2012-2015 Snowplow Analytics Ltd