Druid pour l'analyse de données en temps réel

Yann Esposito

7 Avril 2016

Abstract

Druid expliqué rapidement, pourquoi, comment.

Contents

1 Intro

1.1 Plan

- Introduction ; pourquoi ?
- Comment ?

1.2 Expérience

- Real Time Social Media Analytics

1.3 Demande

- Twitter: `20k msg/s`, `1msg = 10ko` pendant 24h
- Facebook public: 1000 à 2000 msg/s en continu

1.4 En pratique

- Twitter: 400 msg/s en continu, pics à 1500

1.5 Origine (PHP)

1.6 Introduction

- Traitement de donnée gros volume + faible latence
- Typiquement `pulse`

DEMO

2 Druid

2.1 Who

Metamarkets

2.2 Goal

> Druid is an open source store designed for real-time exploratory analytics on large data sets.
>
> hosted dashboard that would allow users to arbitrarily explore and visualize event streams.

2.3 Concepts

- Column-oriented storage layout
- distributed, shared-nothing architecture
- advanced indexing structure

2.4 Features

- fast aggregations
- flexible filters
- low latency data ingestion

arbitrary exploration of billion-row tables tables with sub-second latencies

2.5 Proof