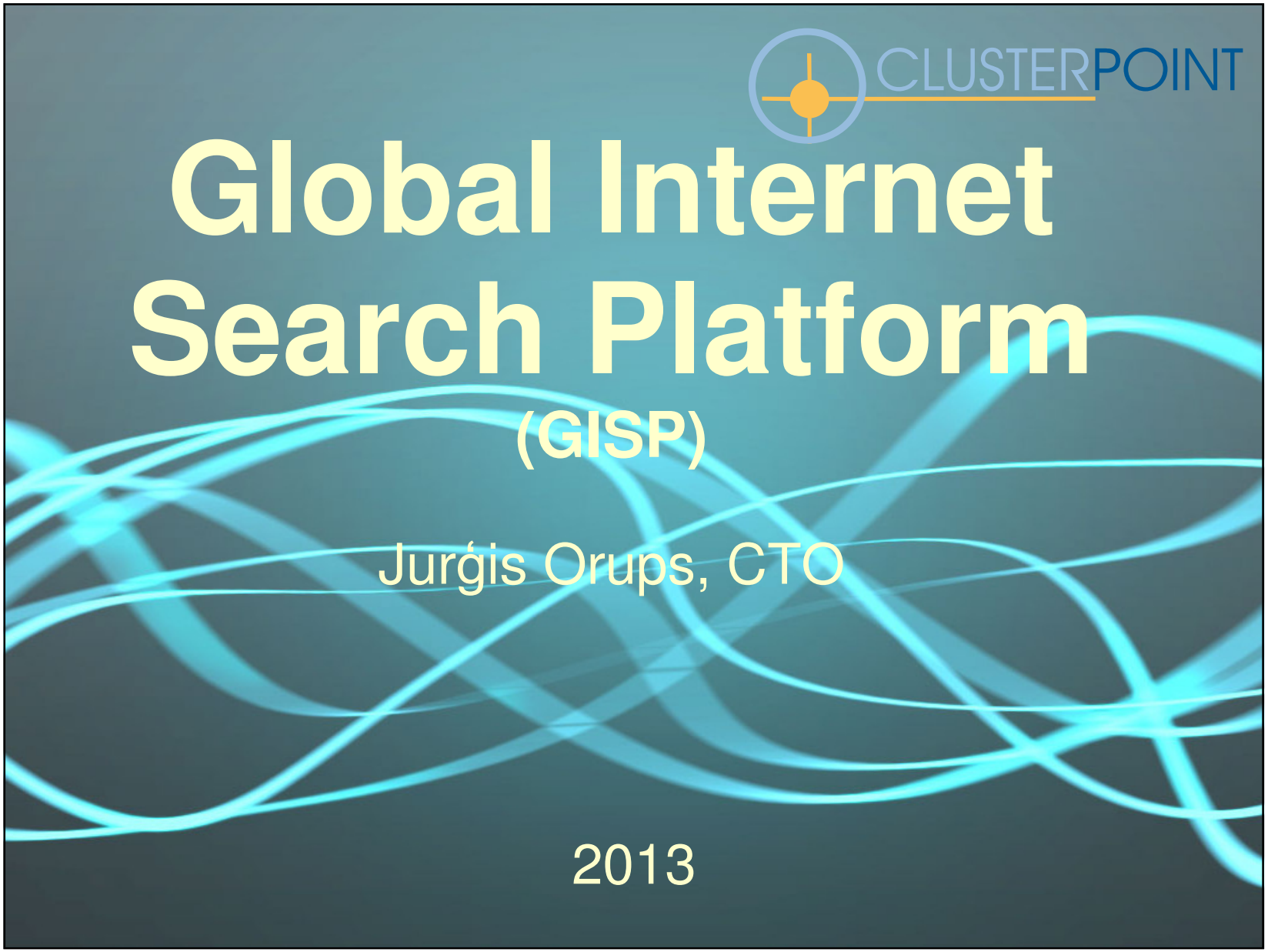
CLUSTERPOINT
Global Internet Search Platform
(GISP)
Jurģis Orups, CTO
2013

Free Google Search API: Gone?

1) Google's Search API is deprecated since Nov 1, 2010. https://developers.google.com/web-search/

2) Google's Deprecation Policy states that they are going to run the service for three years, that is until Nov 1, 2013. https://developers.google.com/web-search/terms

3) As a replacement to Google Search API, Google has Custom Search API, which is a paid service as soon as you go over 100 queries per day. https://developers.google.com/custom-search/v1/overview

4) Official published pricing has tiers upto $2,000 / 500,000 queries.

5) Unofficially there are reports of $200,000 + annually asked for the load to be about 300,000 queries per day at the time.

6) Lesson: big companies trading on stock market can not afford unlimited free services for long term

7) Content aggregated by big companies will eventually be pay-walled against competing businesses

Own Content Aggregation?

- Your own database is fully controlled for your business
- Unrestricted data use vs some cloud API terms of use
- Your partners can add more value with quality content
- Long-term business budgeting at fixed OPEX costs
- Your business does not depend on cloud uptime/outage
- Local ranking vs one-size-fits-all page rank for relevance
- No spam in the controlled content environment
- You keep entire IP vs regular subscription to a cloud API
- Better ROI after 6-9 months, than using Amazon-cloud
- No vendor locked-APIs (no cost escalation risk later)

Outline

- Overview
- Crawler features
- Content processing & indexing
- Search
- Use case
- Q&A

What is GISP?

Scalable, distributed platform for
Web crawling & information retrieval

Based on Clusterpoint DBMS

Process flow

1) **Seed** - enter URLs to start from
2) **Fetch** – collects data from web pages
3) **Process –** parses content
4) **Update** – appends new URL to list
5) **Index –** stores & index data in CP DBMS
6) Return to 2) until all pages visited

Outline

- Overview
- **Crawler features**
- Content processing & indexing
- Search
- Use case
- Q&A

Features

- Tasks
  - Fetch data from web
  - Collect links and follow them
  - Parse content (find title, body, link names, images, etc.)
  - Process binary formats for indexing
  - Identifies content language (en, rus, lv)

Features cont.

- Architecture
  - Clustered
  - Multi-threaded
- Follows robots.txt
- Keeps track of changes in pages
- Protocols
  - http, https, ftp, file, ncp

Features cont.

- Crawling strategy
  - Depth-first
  - Breadth-first
- IP throttling
- Limit URLs by patterns
- Authorization
  - HTTP
  - FTP
  - SMB

Features cont.

- Page ranking
  - Incoming link count & rank
  - URL depth
  - Boosting
  - Refresh rate
- Duplicates
  - Filtering domain aliases

Features cont.

- Configuration options
  - Fetch interval per domain
  - Max pages/depth/size per domain
  - Parallel domains per cluster node
  - URL patterns
  - Bandwidth throttling
  - Original content cache

Features cont.

- XML API
  - Create/remove crawler tasks
  - Start/stop tasks
  - Add/remove domain
  - Crawl monitoring
  - ~20 API commands

Outline

- Overview
- Crawler features
- **Content processing & indexing**
- Search
- Use case
- Q&A

Content processing

- Formats
  - html, doc, ppt, xls, pdf, ps, rtf, eml
  - arj, gz, rar, tar, zip
- Identifies language of content
  - latvian, russian, english
- Programmable patterns
  - Alerting (lua)
  - Filtering (xsl stylesheets)

Content processing cont.

- Stored & indexed by Clusterpoint DBMS
  - Document oriented
  - Scalable
  - Fast
  - Provides full text search functionality
  - Customizable search result ranking
  - Easy integration through rich API

Outline

- Overview
- Crawler features
- Content processing & indexing
- **Data retrieval**
- Use case
- Q&A

Data retrieval

- Search
  - Ad-hoc queries
  - Contextual search
  - Wildcard/pattern search
  - Faceted search
  - Geospatial search
  - Stemming, similarity, dictionaries
- Data aggregation
  - Statistics & reporting

Outline

- Overview
- Crawler features
- Content processing & indexing
- Data retrieval
- **Use case**
- Q&A

Use case

- Latvian Internet search system
  - Servers: 8 (32GB RAM/2TB HDD)
  - Domain count: ~200k
  - Max pages from domain: 3000
  - Fetch interval: 7sec
  - Simultaneous domains: 1200
  - Time for initial/next crawls: 2weeks/1week
  - Total pages: 30M

Outline

- Overview
- Crawler features
- Content processing & indexing
- Data retrieval
- Use case
- **Q&A**

Q&A