Enabling Geospatial Analytics on Unstructured Big Data

Charlie Greenbacker
Geo DC – 6 Feb 2012

What is CLAVIN?

Cartographic Location And Vicinity INdexer

Geoparser

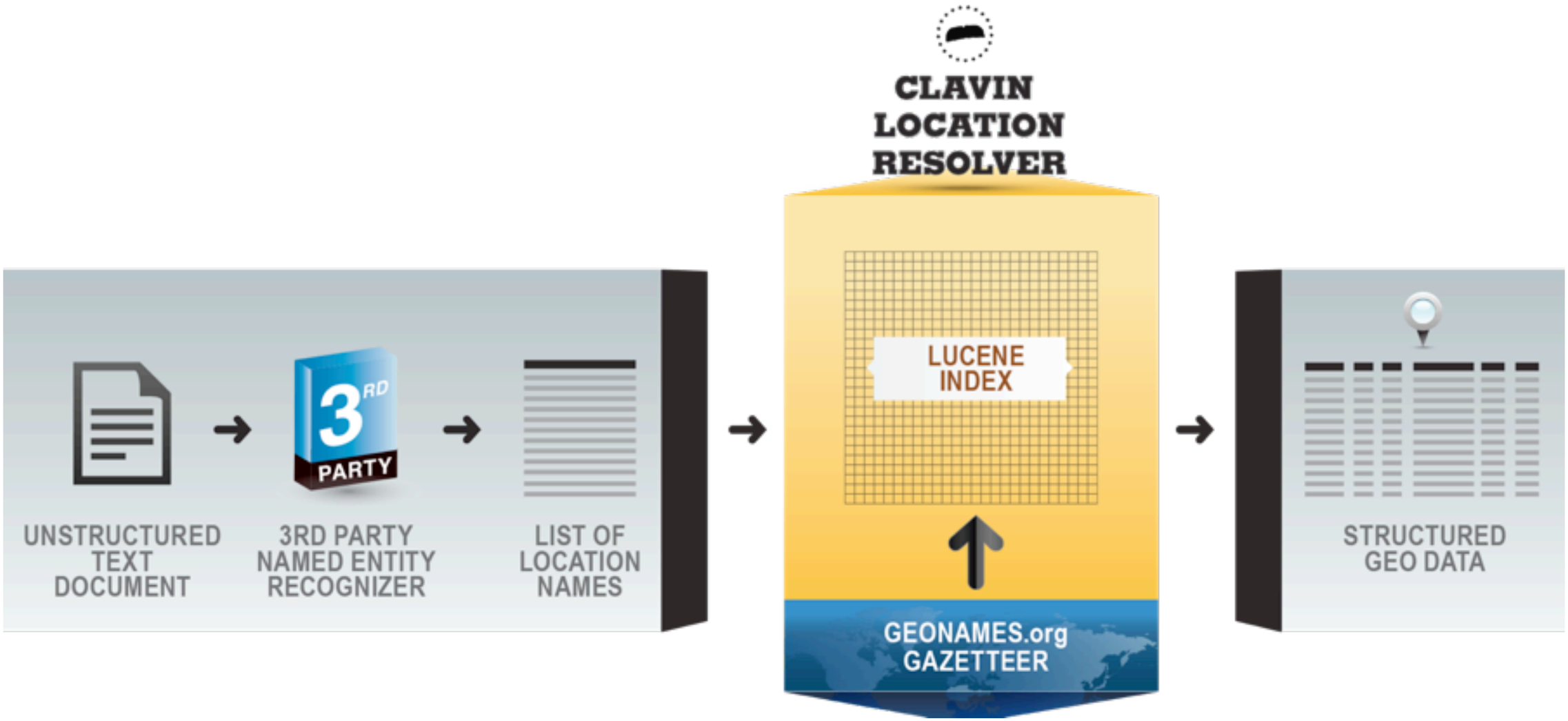
Extracts location names

Resolves geospatial entities

Open source

Runs on Hadoop

Under the hood



CLAVIN handles

Ambiguous references

Typos & phonetic spellings

"Ivory Coast" == "Côte d'Ivoire"

Demo Time!

Only the top 20 locations are shown.

Locations Parsed and Resolved From Text

ID	Name	Lat, Lon	Country Code	#
6446345	US	49.1, 1.96667	FR	9
298795	Turkey	39.05901, 34.91155	TR	9
2215636	Libya	28, 17	LY	8
88319	Benghazi	32.11667, 20.06667	LY	7
2215636	Libyan	28, 17	LY	5
3199389	Herzegovina	43, 17.83333	BA	1
1149361	Afghanistan	33, 66	AF	1
529468	Southeast Europe	43.5333, 3.9833	FR	1
3319	Bengazi	32.11667, 20.06667	LY	1
0630	Cairo	30.06263, 31.24967	EG	1
5455014	American	33.52813, -105.74332	US	1
831053	Kosovo	42.58333, 21	XK	1
3277605	Bosnia	44.25, 17.83333	BA	1
685608	Balkans	44, 23	RO	1
6783140	Gadaffi	12.43269, 14.24894	CM	1
783754	Albania	41, 20	AL	1
718075	Macedonia			

CLAVIN stats

Accurate: 0.75 f-measure

Fast: 100 locations per second per CPU

Scalable: processes 1 million documents in under 1 hour on a 9-node Hadoop cluster

Free: zero licensing cost (Apache License)

clavin.bericotechnologies.com

@CLAVIN__ (two underscores)
@greenbacker
@BericoTech (we're hiring!)