Introduce Spark

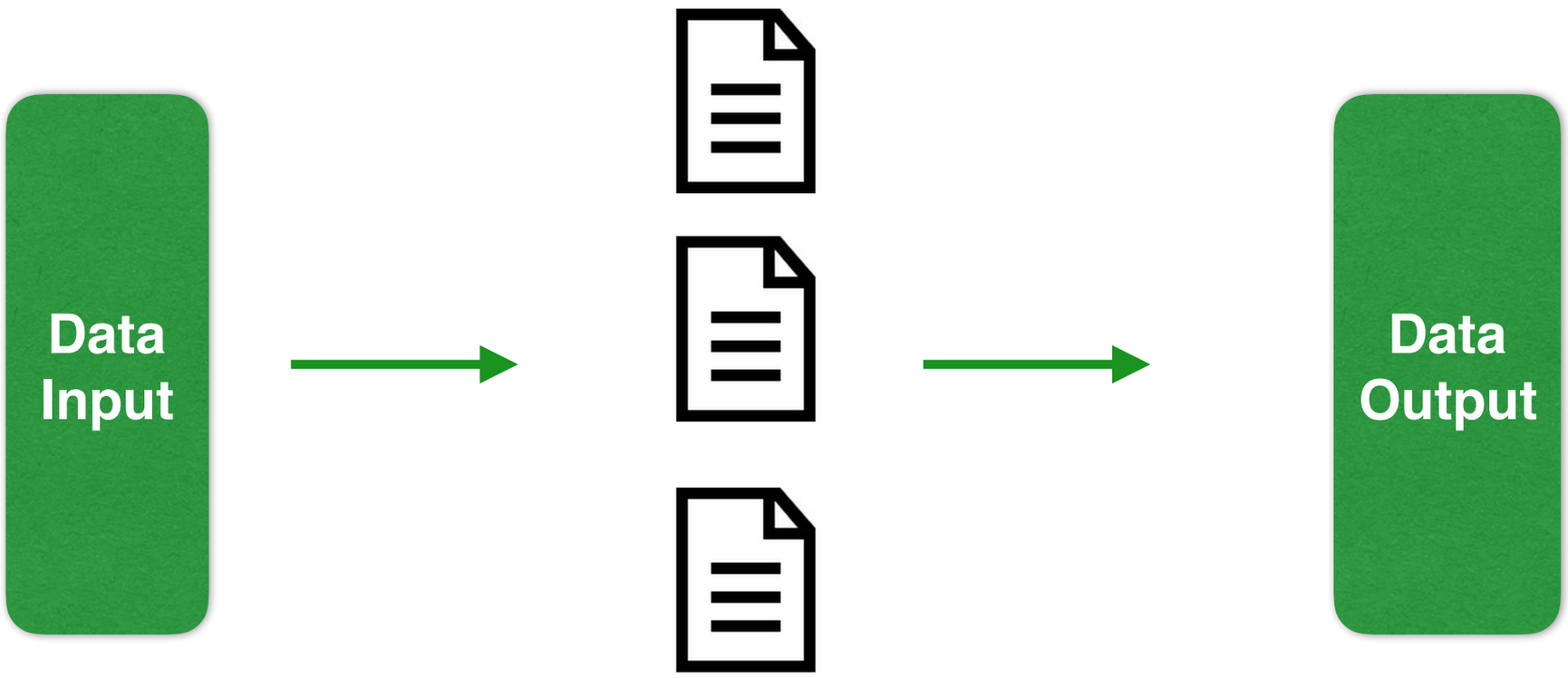
Hanhan Wu

Faster, Easier & Smarter

Agenda

- Faster
- Easier
- Smarter
- Spark Cloud Demo

Faster

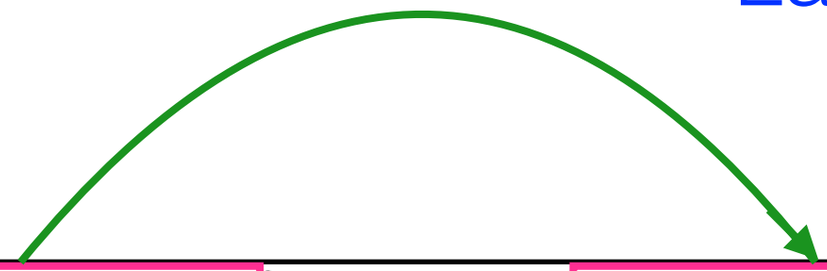


Same code will be executed on different units

in parallel

Faster

Location is calculated on each
Latitude & Longitude
in parallel



id	latitude	longitude	post_date	level1_locations
25652783303	51.689013	5.303413	2016-04-05	's-Hertogenbosch, Noord-Brabant, Netherlands
25652737943	51.392350	-116.412506	2016-04-05	Field, British Columbia, Canada
25652465823	64.760038	-23.880286	2016-04-05	West, Iceland
25650224804	11.162908	119.397869	2016-04-05	El Nido, MIMAROPA, Philippines
26254845555	37.627318	-112.160160	2016-04-05	Bryce, Utah, United States
25981959830	49.446538	11.146455	2016-04-05	Nürnberg, Bayern, Germany

Easier

Spark Pipeline is just a few lines of code

Data Input

"<xml>Stanford University is located in California.
It is a great university.</xml>"

Spark Pipeline

Clean XML tags

Split text into sentences

Just use “.” to add new steps

```
val output = input
  .select(cleanxml('text).as('doc))
  .select(explode(ssplit('doc)).as('sen))
  .select('sen, tokenize('sen).as('words), ner('sen).as('nerTags), sentiment('sen).as('sentiment))
```

Tokenize sentences

Generate Name Entity tags

Sentiment analysis each sentence

Easier

Data Input

"<xml>Stanford University is located in California.
It is a great university.</xml>"

Data Output

```
+-----------------------------------------+-----------------------------------------------------+
|sen                                      |words                                                |
+-----------------------------------------+-----------------------------------------------------+
|Stanford University is located in California .|[Stanford, University, is, located, in, California, .]|
|It is a great university .               |[It, is, a, great, university, .]                    |
+-----------------------------------------+-----------------------------------------------------+
```

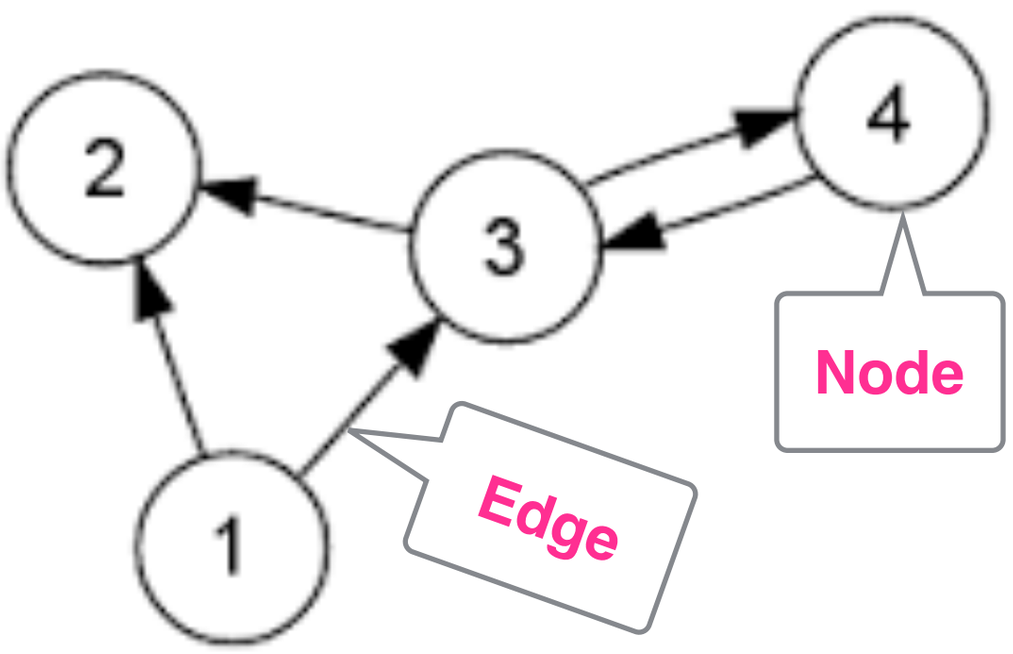
```
+-----------------------------------------------------+---------+
|nerTags                                              |sentiment|
+-----------------------------------------------------+---------+
|[ORGANIZATION, ORGANIZATION, O, O, O, LOCATION, O]|1        |
|[O, O, O, O, O, O]                                   |4        |
+-----------------------------------------------------+---------+
```

Smarter

- **Spark SQL, Dataframe** - Convenient data operations
- **MILib** - Machine Learning Algorithms
- **GraphFrames** - Graph Computation
- **Streaming** - Real time analysis on streaming data
- **Open Source Support**

Smarter

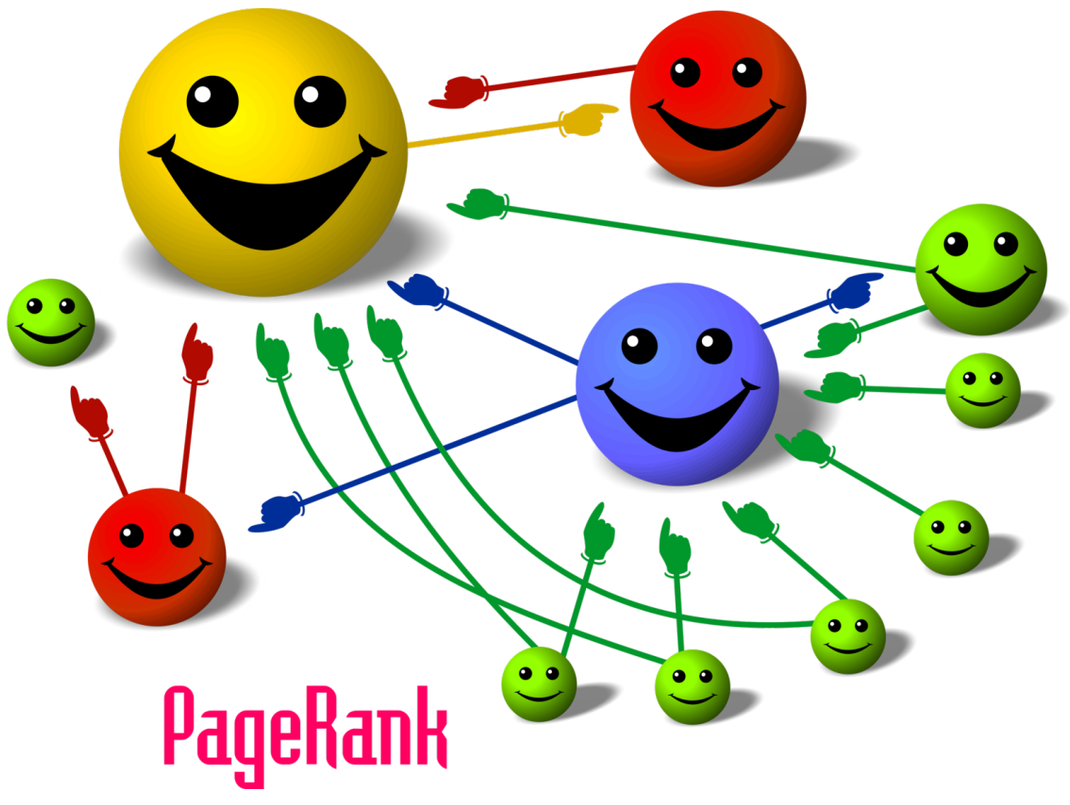
Spark Graph Frames makes graph calculation simple & fast



A graph contains
nodes and edges

Smarter

Page Rank measures how popular a node is



Page Rank

- Urls point to each other
- More reference, more popular
- Graph Node is the url
- Graph Edge is the reference

Smarter

Mockup Data Input

Source	Destination	relationship
url_id1	url_id2	contains
url_id2	url_id3	contains
url_id3	url_id2	contains
url_id5	url_id3	contains
url_id4	url_id5	contains
url_id4	url_id6	contains
url_id6	url_id1	contains
url_id1	url_id5	contains

Smarter

Spark Graph Frames handles all Page Rank calculation

Real World Data Output

id	url	pagerank
579	https://en.wikipedia.org/wiki/Denmark	0.15607142857142856
580	https://en.wikipedia.org/wiki/Smart_phones	0.15607142857142856
581	https://en.wikipedia.org/wiki/Upper_class	0.15607142857142856
582	https://en.wikipedia.org/wiki/Backpack#Backpacks_for_outdoor_activities	0.15607142857142856
584	https://en.wikipedia.org/wiki/Hostelling_International	0.15607142857142856
585	https://en.wikipedia.org/wiki/Authenticity_(philosophy)	0.15607142857142856
583	https://en.wikipedia.org/wiki/Hippie_trail	0.15607142857142856
292	https://en.wikipedia.org/wiki/Lifestyle_travelling	0.15022135416666665
492	https://en.wikipedia.org/wiki/Sex_tourism	0.15022135416666665
93	https://en.wikipedia.org/wiki/Vagrancy_(people)	0.15022135416666665
293	https://en.wikipedia.org/wiki/Amikeca_Reto	0.15022135416666665
493	https://en.wikipedia.org/wiki/Port_of_Szczecin	0.15022135416666665

Spark Cloud Demo

Great Documentation

Spark 2.1.0 Documentation

Spark Python

Spark Cloud Login

Spark Cloud User Guide