# Sentiment Analysis

in

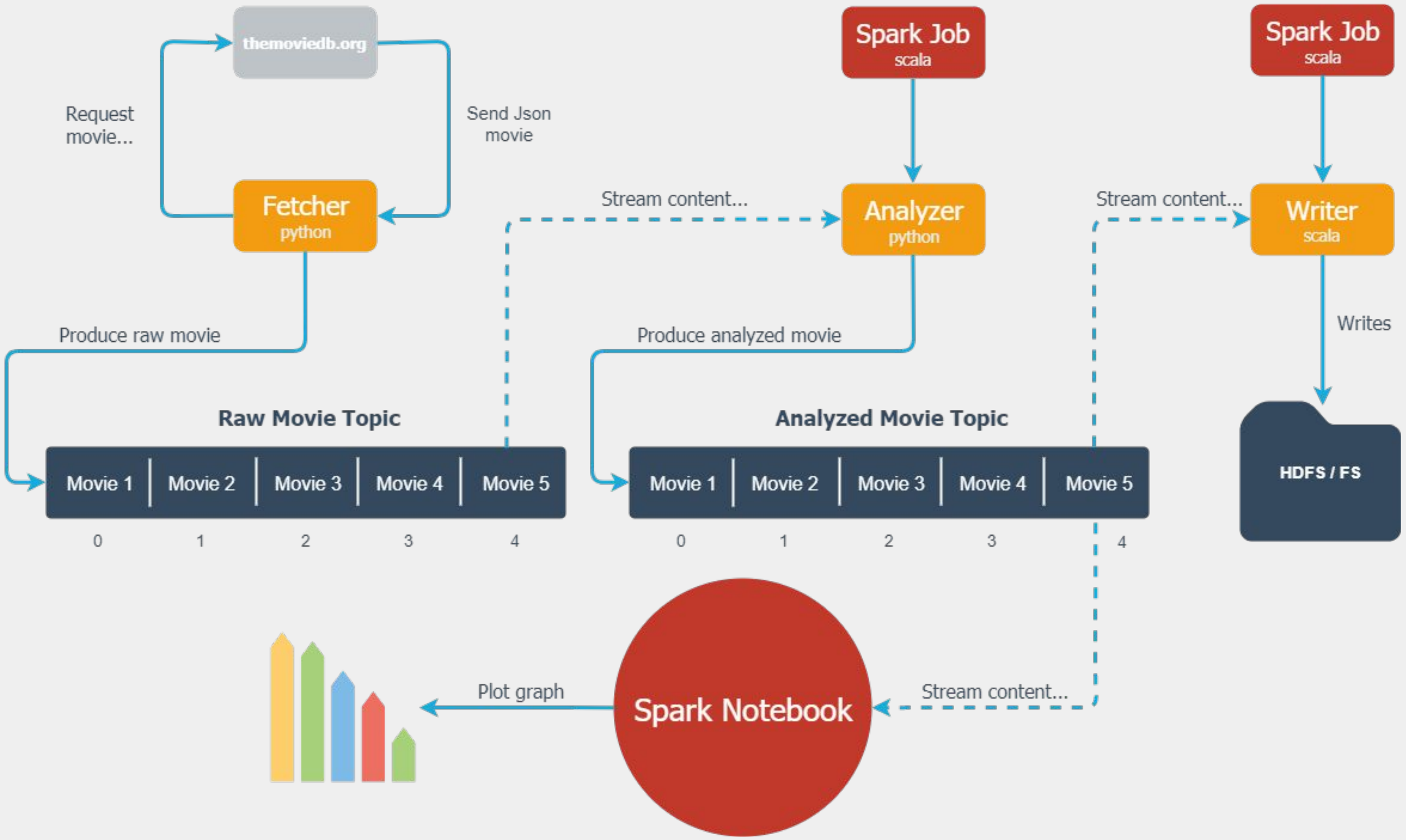
APACHE Spark™

Sarasvati Moutoucomarapoulé
David Peicho

# Pipeline

- Stack Overview
- Python Fetcher
- Sentiment Analysis
- Notebook

# Stack Overview



**Raw Movie Topic**

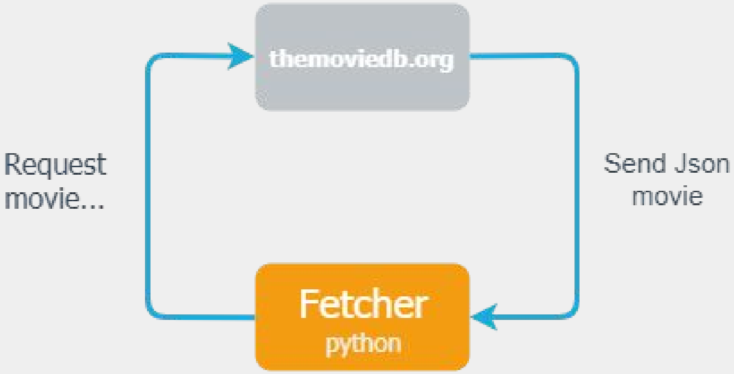
| Movie 1 | Movie 2 | Movie 3 | Movie 4 | Movie 5 |
|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 |

**Analyzed Movie Topic**

| Movie 1 | Movie 2 | Movie 3 | Movie 4 | Movie 5 |
|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 |

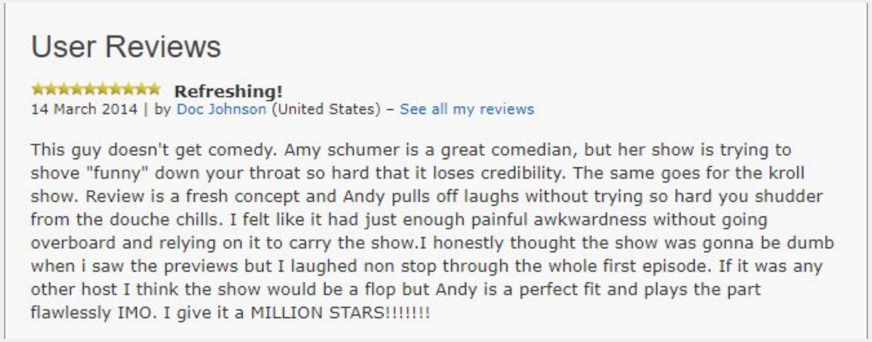
# Python Fetcher

- Works as a service
- Fetches data from https://www.themoviedb.org/
- Fetches multiple data:
  - Title
  - Release date
  - Popularity
  - …
- Produces to a specific topic



# Sentiment Analysis: Python Script

- Trained on Large Review Dataset V1.0
  - 25.000 positive reviews
  - 25.000 negative reviews
- Based on a Linear SVM
  - Input: matrix of TF-IDF features
- **85%** of good classification
- Produces to a specific topic



User Reviews

★★★★★★★★★★ **Refreshing!**

14 March 2014 | by Doc Johnson (United States) – See all my reviews

This guy doesn't get comedy. Amy schumer is a great comedian, but her show is trying to shove "funny" down your throat so hard that it loses credibility. The same goes for the kroll show. Review is a fresh concept and Andy pulls off laughs without trying so hard you shudder from the douche chills. I felt like it had just enough painful awkwardness without going overboard and relying on it to carry the show.I honestly thought the show was gonna be dumb when i saw the previews but I laughed non stop through the whole first episode. If it was any other host I think the show would be a flop but Andy is a perfect fit and plays the part flawlessly IMO. I give it a MILLION STARS!!!!!!!

*User review extracted from imdb.com*

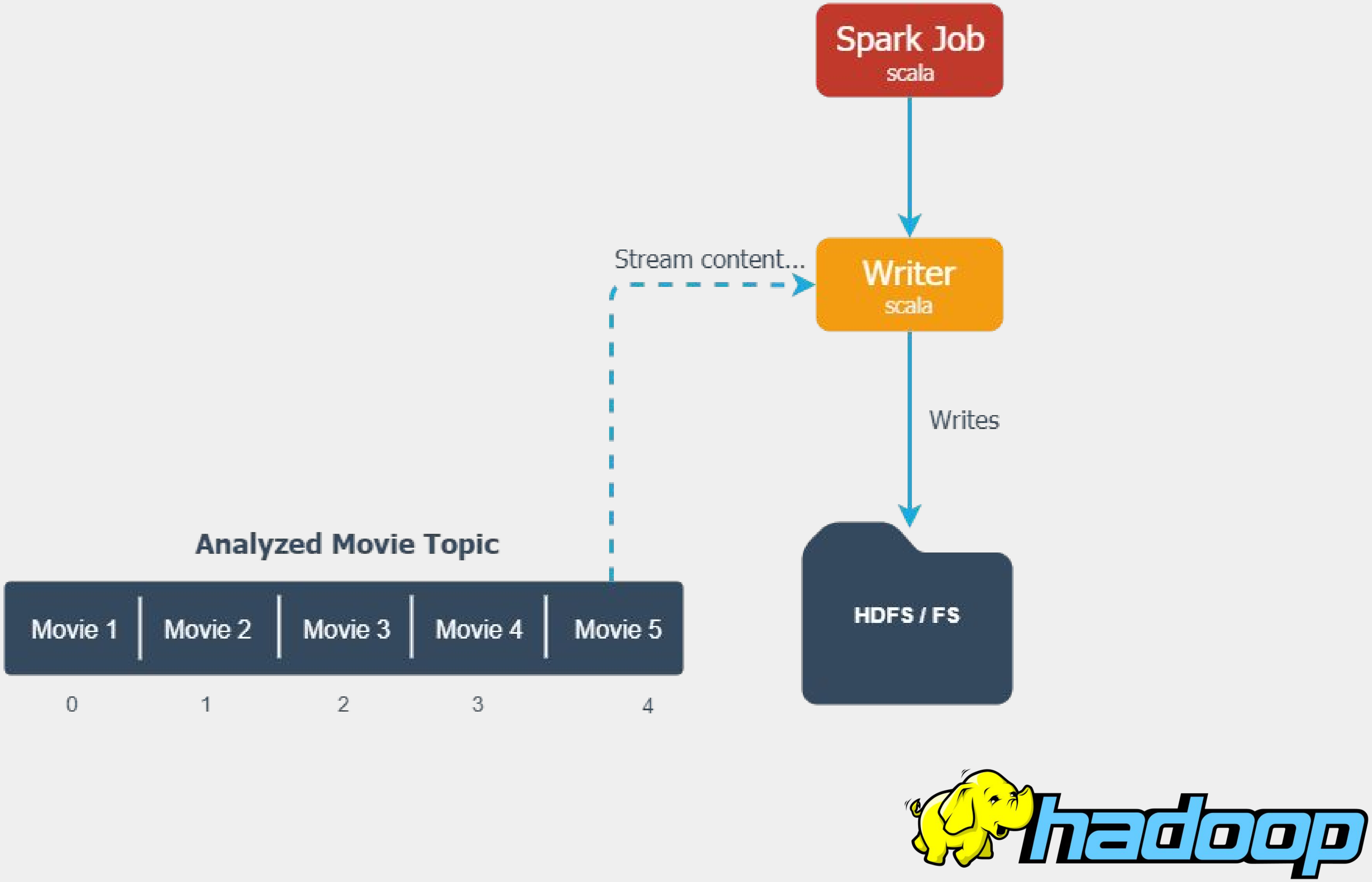
# Sentiment Analysis: Parallelization

- Unique service is not quick enough…

- Solution: use Spark workers

- Analysis applied through the pipe function:
  - Stdin fetches
  - Stdout writes

# Persistence



# Conclusion

# Conclusion

- Themoviedb API has too few reviews
  - Affects analysis

- Sentiment analysis could be improved
  - Naïve Bayes ?

- Spark Notebook not enough documented…

- Kafka and Spark make a good flow!