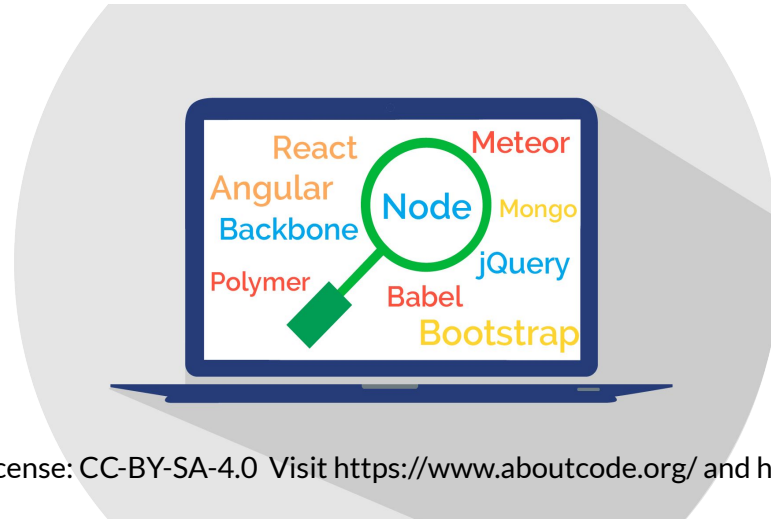


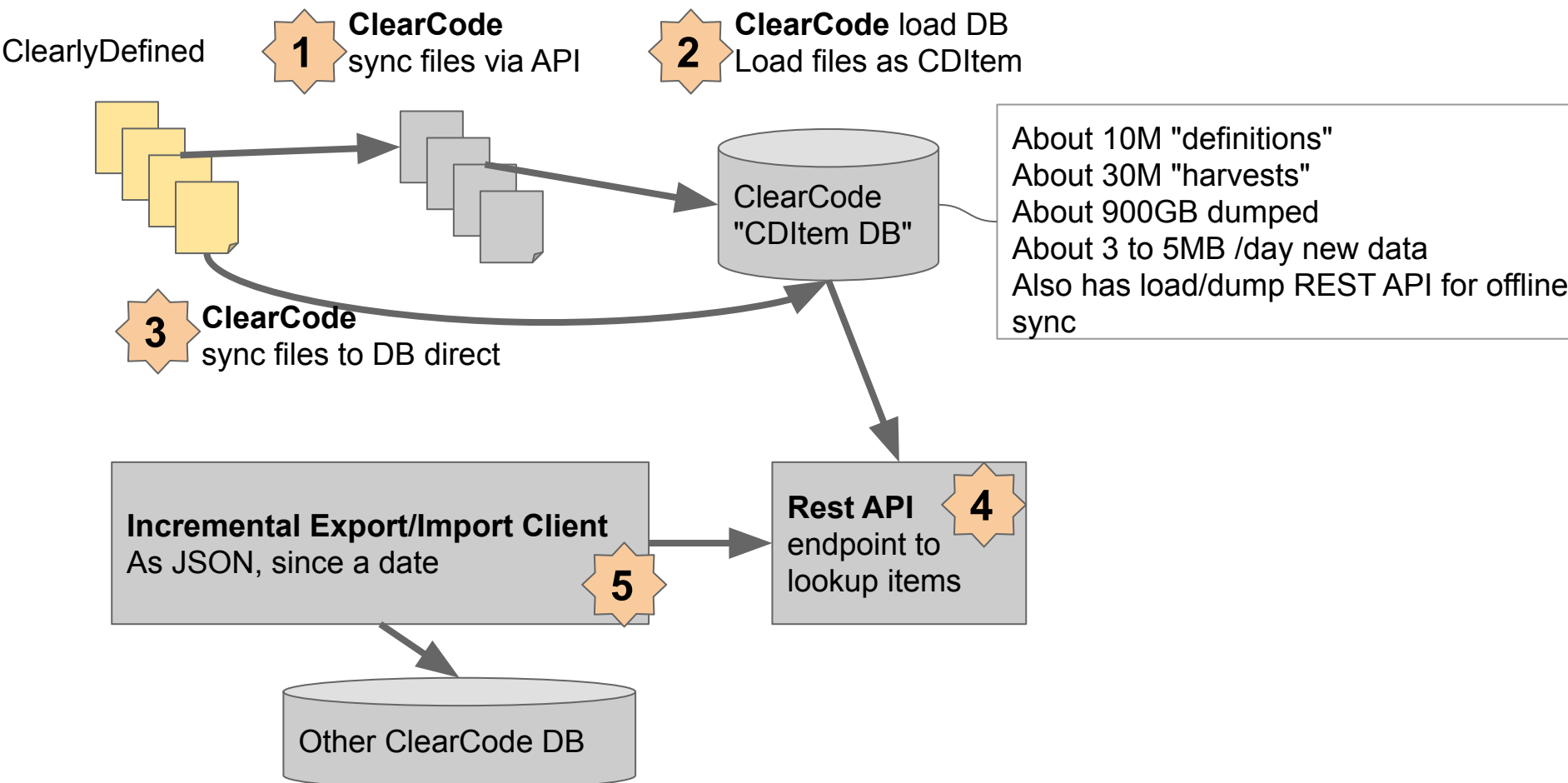
# ClearCode Overview

Summer 2020



# What is ClearCode? Why?

- ▶ **A tool to mirror ClearlyDefined data**
  - And keep it in sync, up-to-the-minute
  - Syncing everything: definitions, harvests, etc.
- ▶ **Why? To use the data wholesale for any use case**
  - Bulk analysis and queries
  - Behind-the-firewall, private usage
  - Integration in various solutions, database, software products.
  - ...



# Tech

- ▷ Python with PostgreSQL
  - Django for ORM, DRF for API, Click for CLI
- ▷ Data stored as original JSON as Gzipped field
- ▷ Record key is the original file path
- ▷ Data can also be synced as plain JSON files
- ▷ Sync is multithreaded and polite by default using etags, and throttling and caching
- ▷ Utilities to convert CD Coordinates to "Package URLs" purls
- ▷ Apache-licensed

# Challenges

- ▷ File storage is impractical outside of Azure
  - ~ 100 million files and dirs... but close to 50 million dirs!
- ▷ Using filesystem paths as a key does not work at that scale
- ▷ Once compressed in a DB, the dataset is much easier to handle
  - less than one TB
- ▷ Cloudflare seems to be a source of many subtle and painful network issues
- ▷ Some ClearlyDefined API have some unstable behaviour which are hard to qualify

# Next Steps

- ▷ Open sourced today at <https://github.com/nexB/clearcode-toolkit/>
- ▷ Robust enough, running continuously for several months to sync up the data
- ▷ We have a whole seed available to share. The challenge is the distribution of ~ 1TB of data
- ▷ This is used as a dataset for a Google Summer of Code project to apply AI/ML to spot inconsistencies and improve ScanCode license detection

# Contact

## ▷ Contact persons

- Michael Herzog  
mjherzog@nxb.com  
+ 1 650 380 0680
- Philippe Ombredanne  
pombredanne@nxb.com  
+ 1 650 799 0949

## ▷ More information

- <https://www.nxb.com/> and <https://www.aboutcode.org/>



# Credits

Special thanks to all the people who made and released these awesome free resources:

- ▷ Presentation template by [SlidesCarnival](#)
- ▷ Photographs by [Unsplash](#)
- ▷ And all the software authors that made ClearCode, AboutCode and ScanCode possible