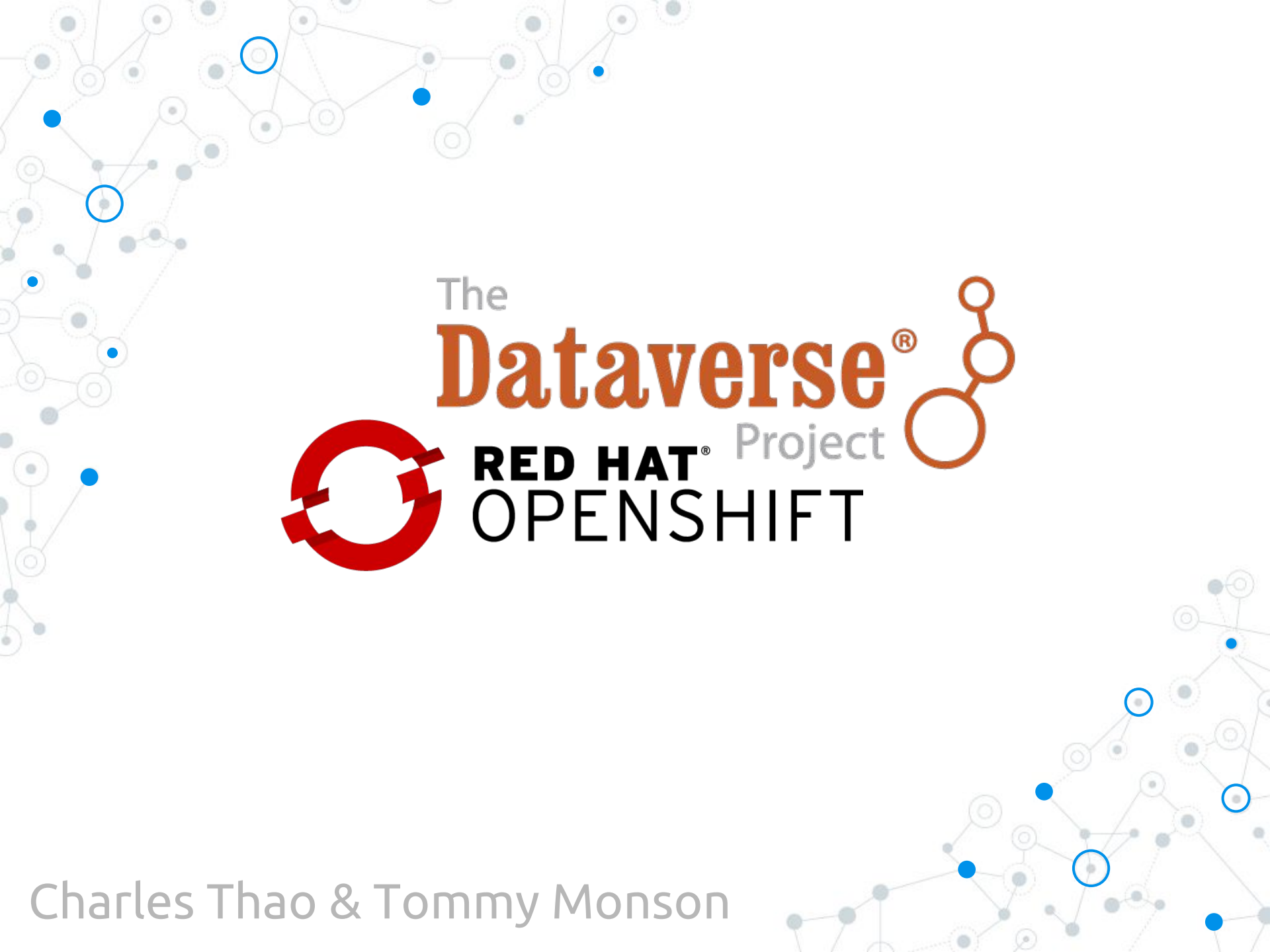


Charles Thao & Tommy Monson

What is Dataverse?

- **An open-source platform to publish, cite, and archive research data**
- Built to support multiple types of data, users, and workflows
- Developed at Harvard's Institute for Quantitative Social Science (IQSS) since 2006
- Hundreds of members of the Dataverse Community - developers, researchers, librarians, data scientists

Dataverse Community

- 33 installations around the world

Why Dataverse?

- DataVerse as a complementary component to DataHub
- A powerful tool for data storage and research

Pod Applications

- Oracle's implementation of Java EE
- Serves as an *application server*
- Supports the Dataverse source code

- Object-relational *database management system*
- Highly scalable
- Primary competitor is mySQL

- Apache's *enterprise search* platform
- Highly scalable and fault-tolerant
- RESTful

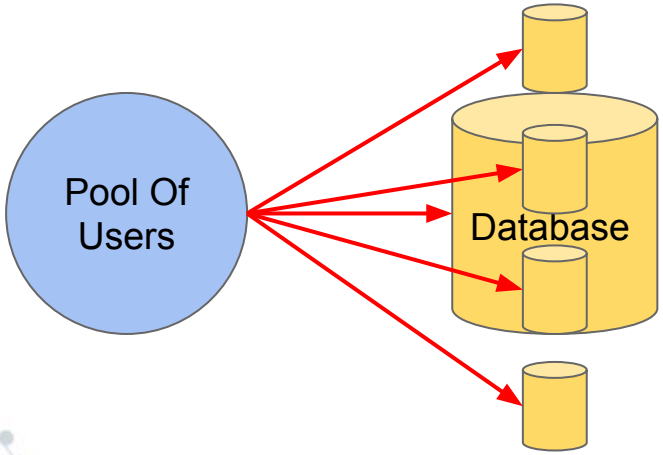
Of course, all of these technologies are open-source.

Hiding Secret Information

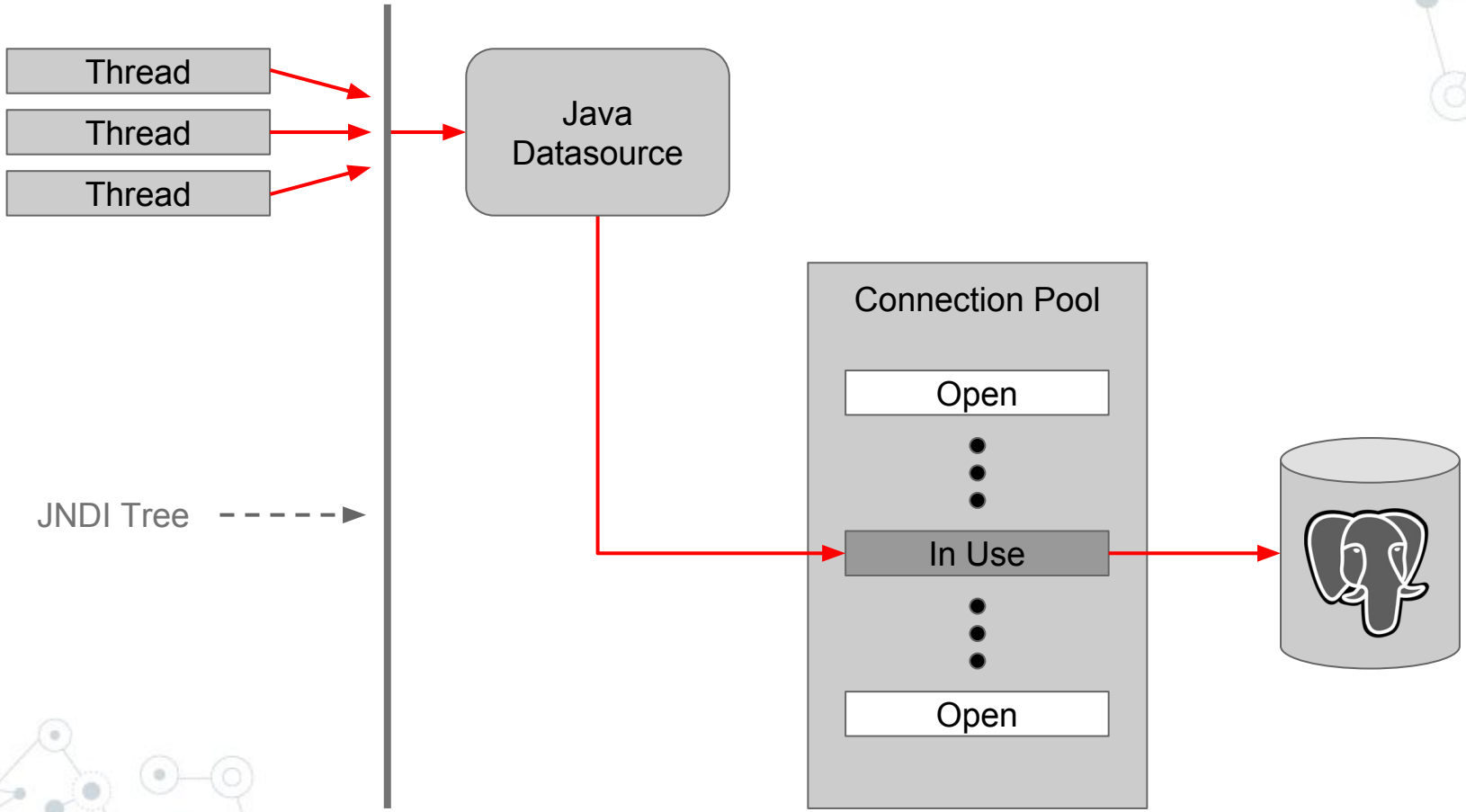
- Writing usernames and passwords explicitly in code is not secure
- Secrets should be randomized and stored in environment variables
- Rooting still exposes the secrets -- cryptographic hashing?

Scaling PostgreSQL

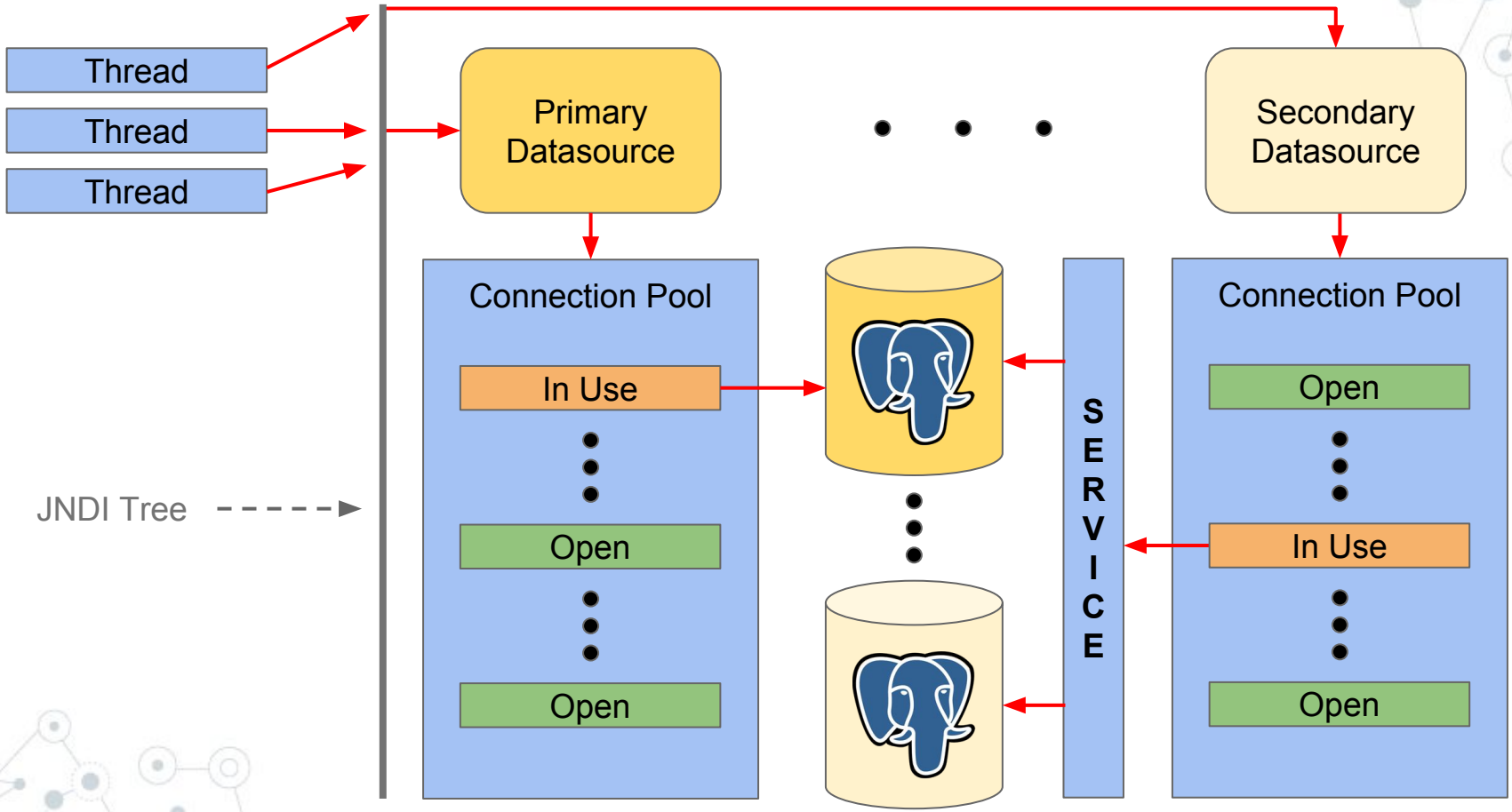
- As the Dataverse project grows, the application will require higher availability
- Implementing multiple databases can achieve this (primary and secondaries)
- Must decide when it is allowable to read from a secondary rather than a primary



PostgreSQL Current State

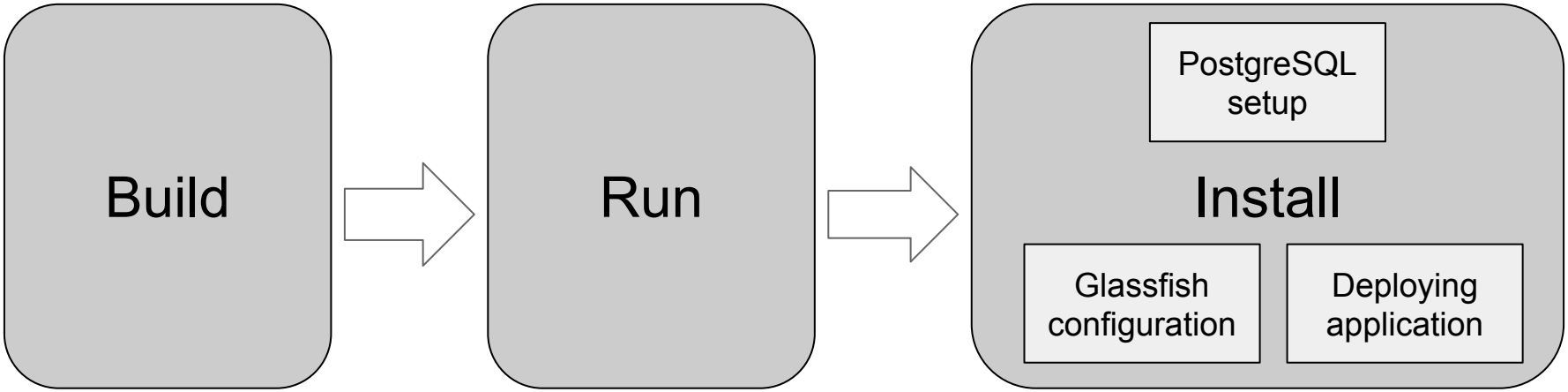


PostgreSQL Future State

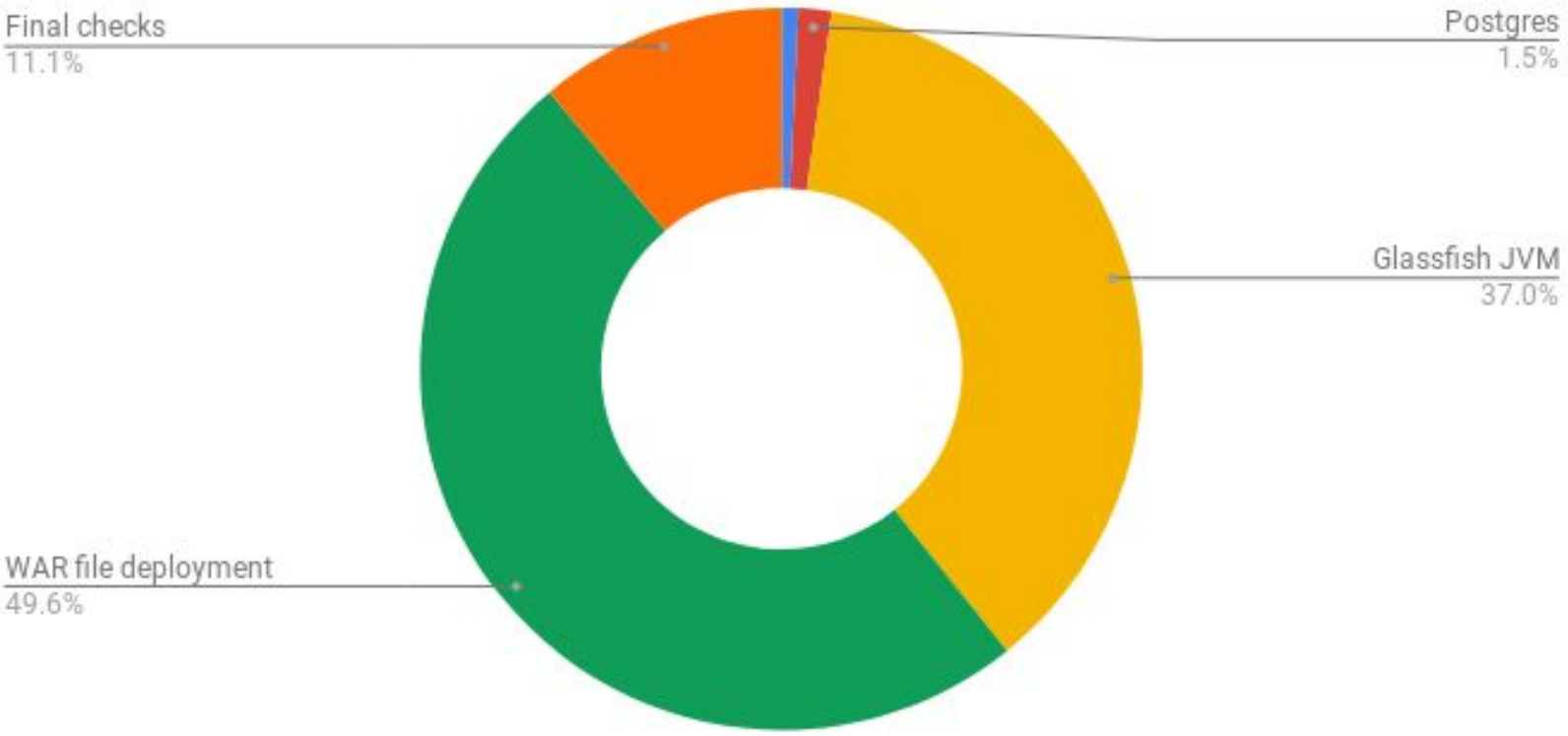


Redesigning Dataverse Installation for Containers

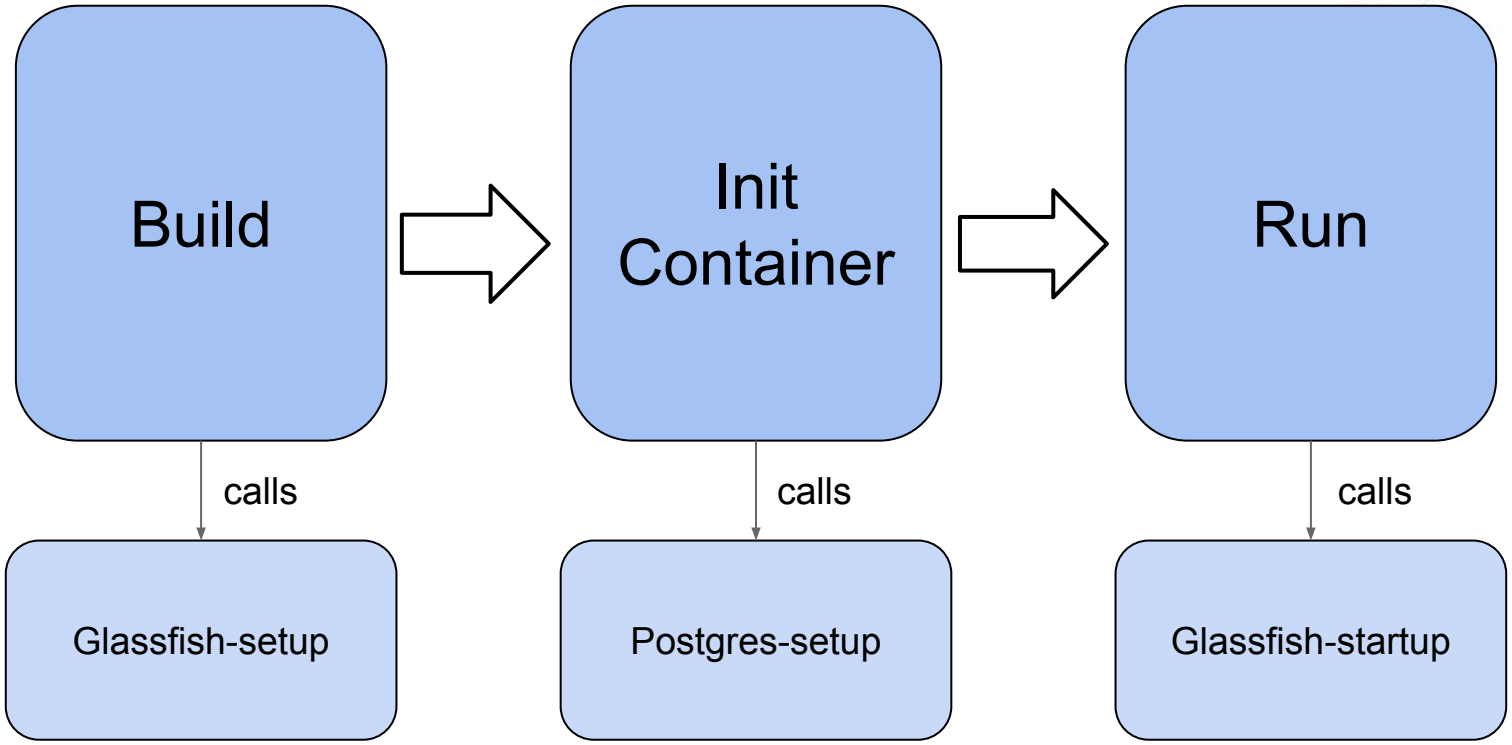
Original Design



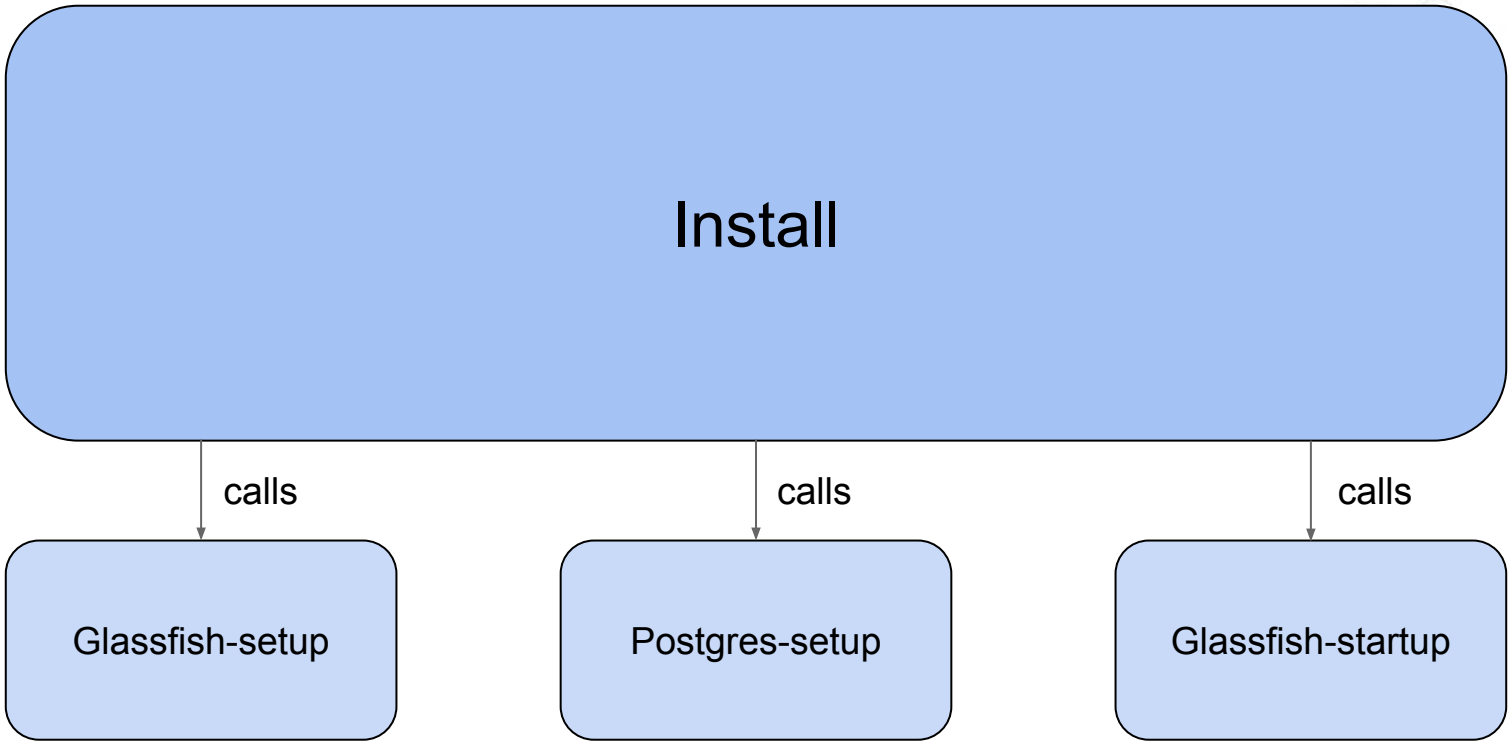
Startup time by parts



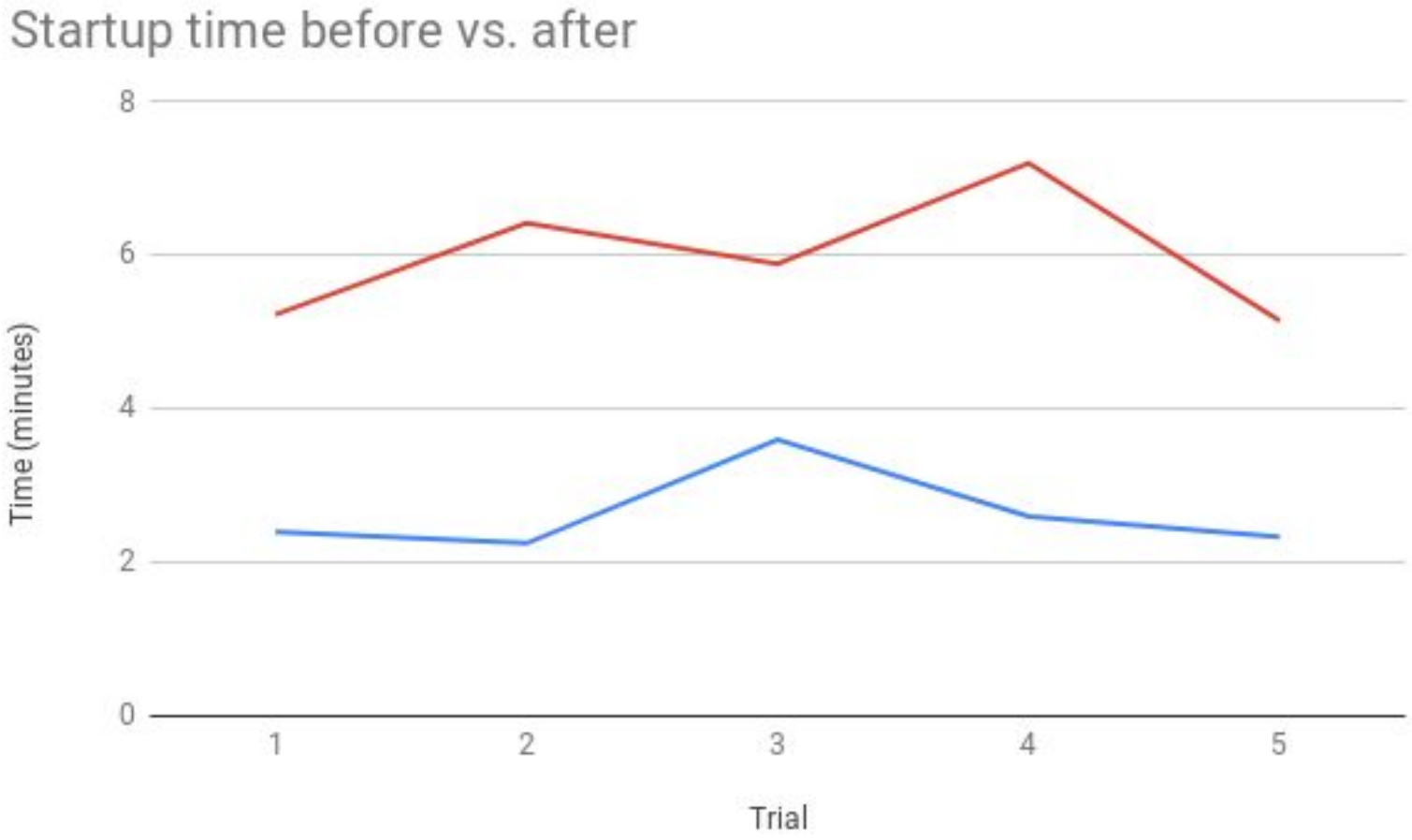
Final Design



For non-container users



Redesigning Dataverse Installation

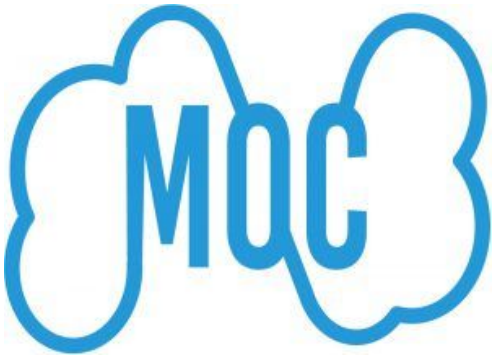


Solr Scaling

- Solr: open source enterprise search platform
- StatefulSets: For stateful applications and distributed systems
  - Previous work in Glassfish and Postgres
- How? Distributed indexing and index replication

Extended goal

Get the DataVerse broker running on UpShift/ MOC




Questions?