Spfy: speedy predictive genomics

Supervisor: Chad Laing
Student: Kevin Le

Kevin Le

Currently:

- BASc Software Engineering (uOttawa)
    - Focus on applied math
- Co-op @NML Lethbridge (8-months)

Previously Completed:

- BSc Neuroscience (Dalhousie)
    - Transgenic mouse models of Alzheimer's
    - Cryptography & network security

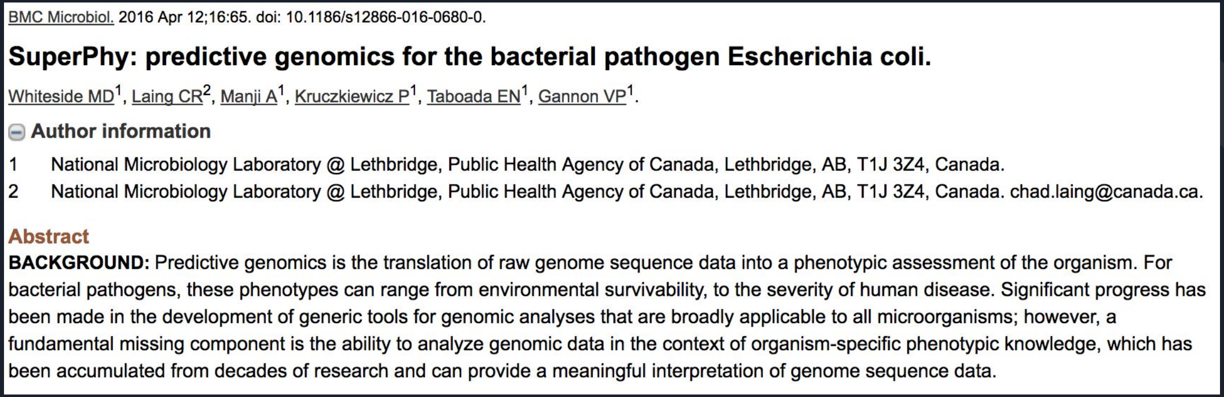
Software Background:

- Largely Python
- Linux
- Virtualization

Career Goals:

- Big-data companies

Background



BMC Microbiol. 2016 Apr 12;16:65. doi: 10.1186/s12866-016-0680-0.

SuperPhy: predictive genomics for the bacterial pathogen Escherichia coli.

Whiteside MD[1], Laing CR[2], Manji A[1], Kruczkiewicz P[1], Taboada EN[1], Gannon VP[1].

⊟ Author information

1 National Microbiology Laboratory @ Lethbridge, Public Health Agency of Canada, Lethbridge, AB, T1J 3Z4, Canada.

2 National Microbiology Laboratory @ Lethbridge, Public Health Agency of Canada, Lethbridge, AB, T1J 3Z4, Canada. chad.laing@canada.ca.

Abstract

BACKGROUND: Predictive genomics is the translation of raw genome sequence data into a phenotypic assessment of the organism. For bacterial pathogens, these phenotypes can range from environmental survivability, to the severity of human disease. Significant progress has been made in the development of generic tools for genomic analyses that are broadly applicable to all microorganisms; however, a fundamental missing component is the ability to analyze genomic data in the context of organism-specific phenotypic knowledge, which has been accumulated from decades of research and can provide a meaningful interpretation of genome sequence data.

The problem

Speed

- Predict serotype, VF, AMR within a few minutes
- More responsive user-interface

Simplicity

- Easier to upload genomes and get results
- Intuitive navigation of the website

Scalability

- Quickly integrate new analyses & results
- Ability to perform group comparisons across thousands of genomes

Challenges deep-dive

Goal: Speed

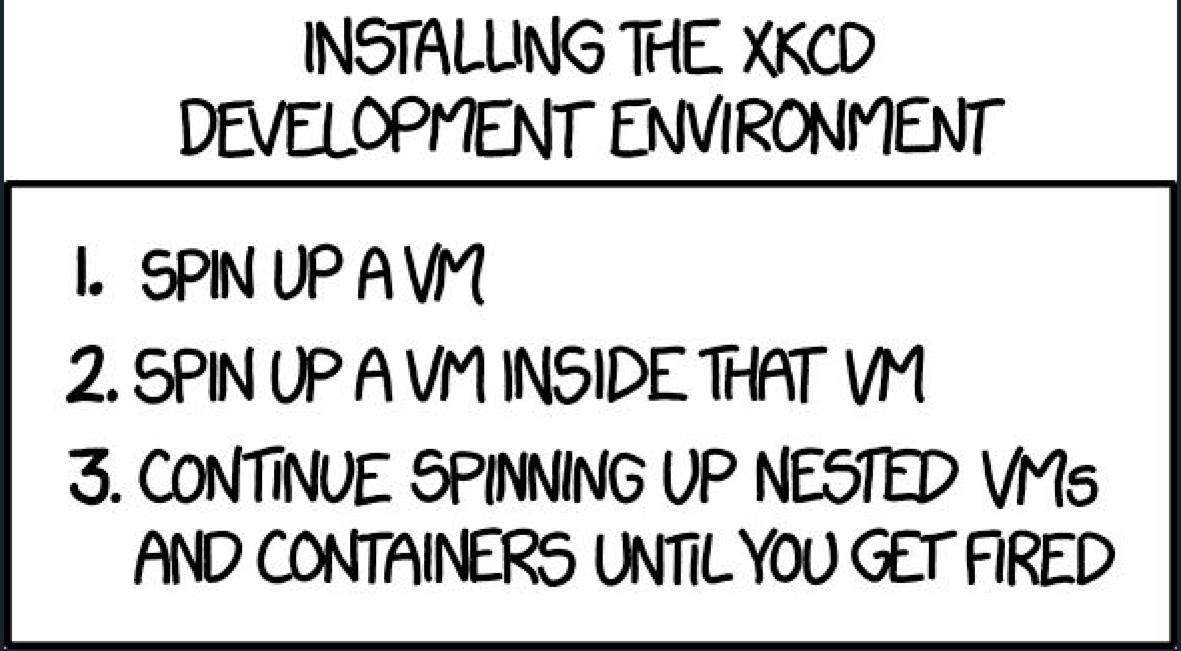
- Predict serotype, VF, AMR within a few minutes
- More responsive user-interface

Solution: Modernize

- Docker
- Task queues
- ReactJS

Docker

Docker

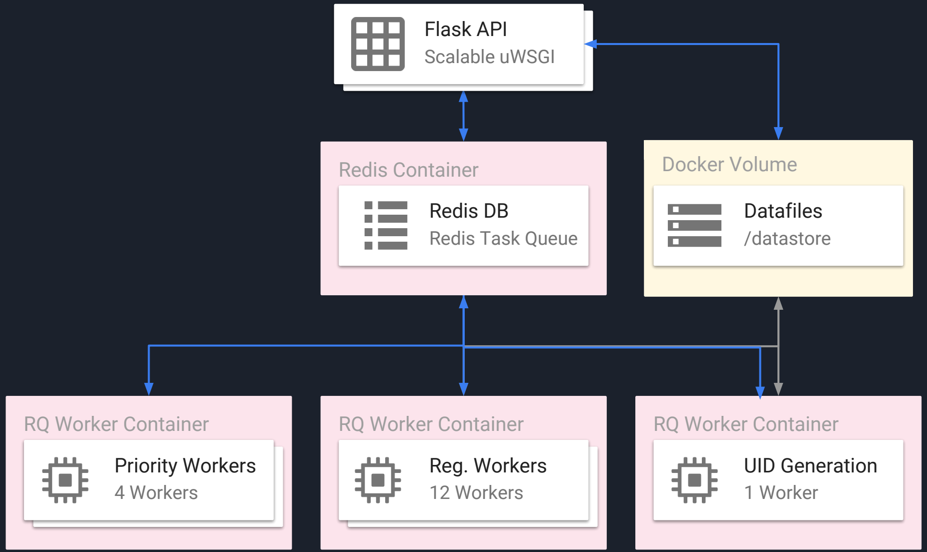


INSTALLING THE XKCD DEVELOPMENT ENVIRONMENT

1. SPIN UP A VM
2. SPIN UP A VM INSIDE THAT VM
3. CONTINUE SPINNING UP NESTED VMs AND CONTAINERS UNTIL YOU GET FIRED

Solution

Speed: task queues



Flask API

Scalable uWSGI

Redis Container

Redis DB

Redis Task Queue

Docker Volume

Datafiles

/datastore

RQ Worker Container

Priority Workers

4 Workers

RQ Worker Container

Reg. Workers

12 Workers

RQ Worker Container

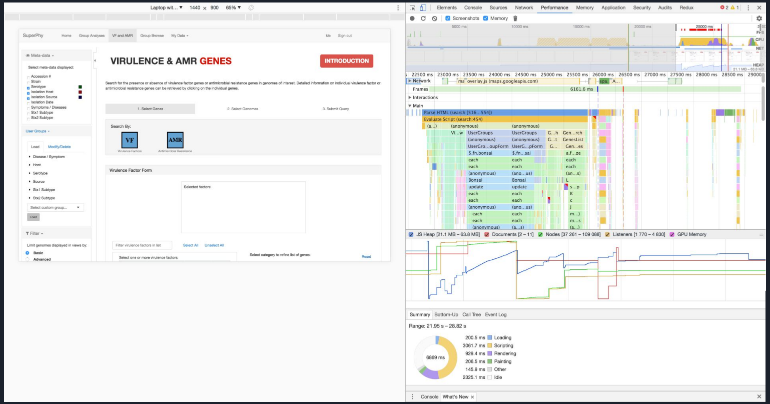
UID Generation

1 Worker

Immediate Results

Solution

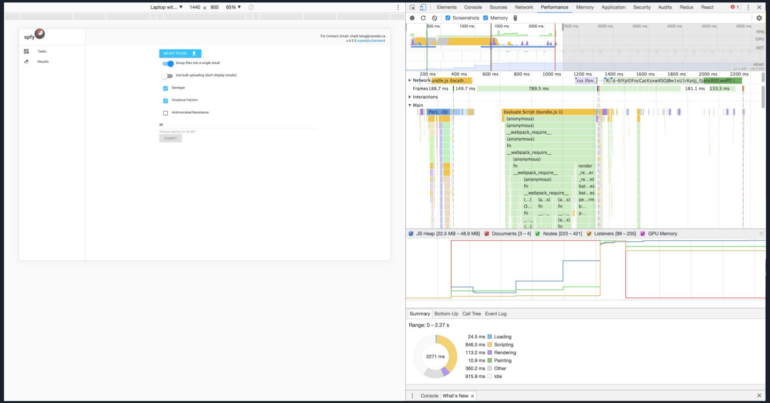
Speed: ReactJS



Range: 21.95 s – 28.82 s

6869 ms

200.5 ms	Loading
3061.7 ms	Scripting
929.4 ms	Rendering
206.5 ms	Painting
145.9 ms	Other
2325.1 ms	Idle



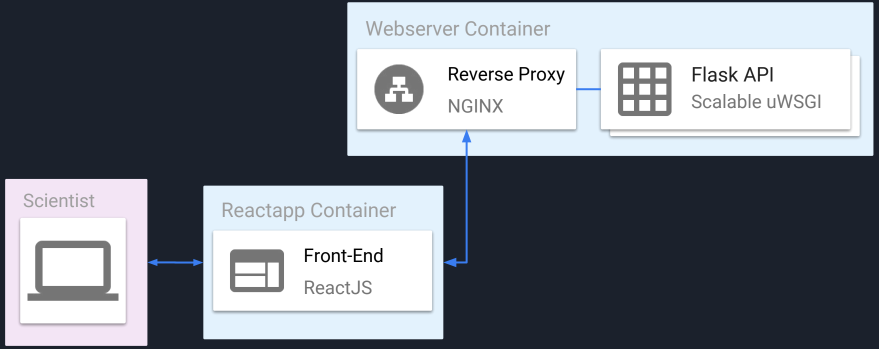
Range: 0 – 2.27 s

2271 ms

24.5 ms	Loading
846.5 ms	Scripting
113.2 ms	Rendering
10.9 ms	Painting
360.2 ms	Other
915.9 ms	Idle

Solution

Speed: ReactJS



Docker

Challenges deep-dive

Goal: Simplicity

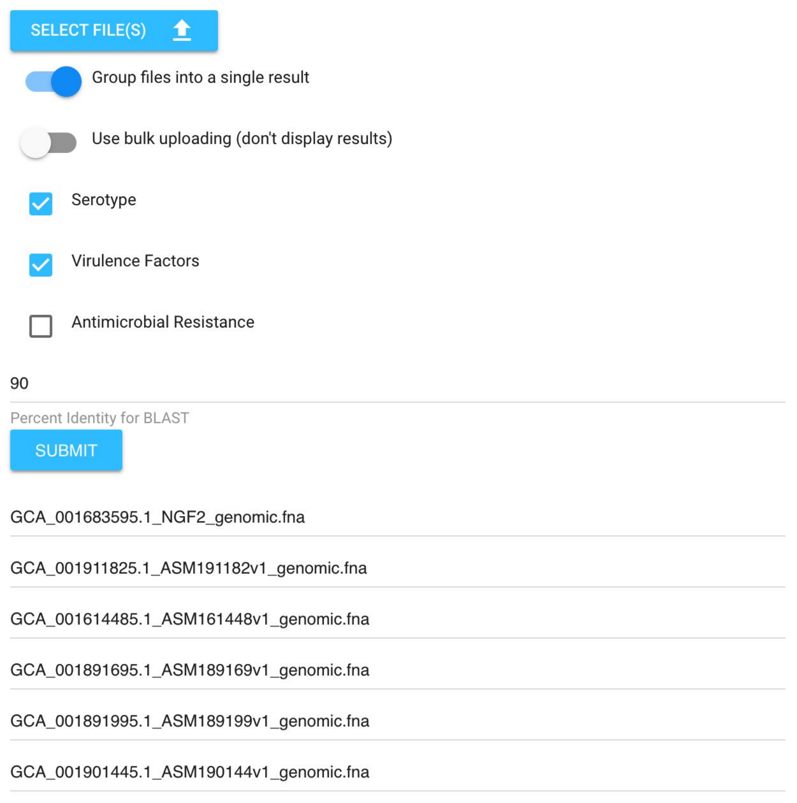
- Easier to upload genomes and get results
- Intuitive navigation of the website

Solution: Familiar Design

- One click results
- Material design (Google)

Solution

Simplicity: one-click results



● Group files into a single result

○ Use bulk uploading (don't display results)

☑ Serotype

☑ Virulence Factors

□ Antimicrobial Resistance

Percent Identity for BLAST: 90

GCA_001683595.1_NGF2_genomic.fna

GCA_001911825.1_ASM191182v1_genomic.fna

GCA_001614485.1_ASM161448v1_genomic.fna

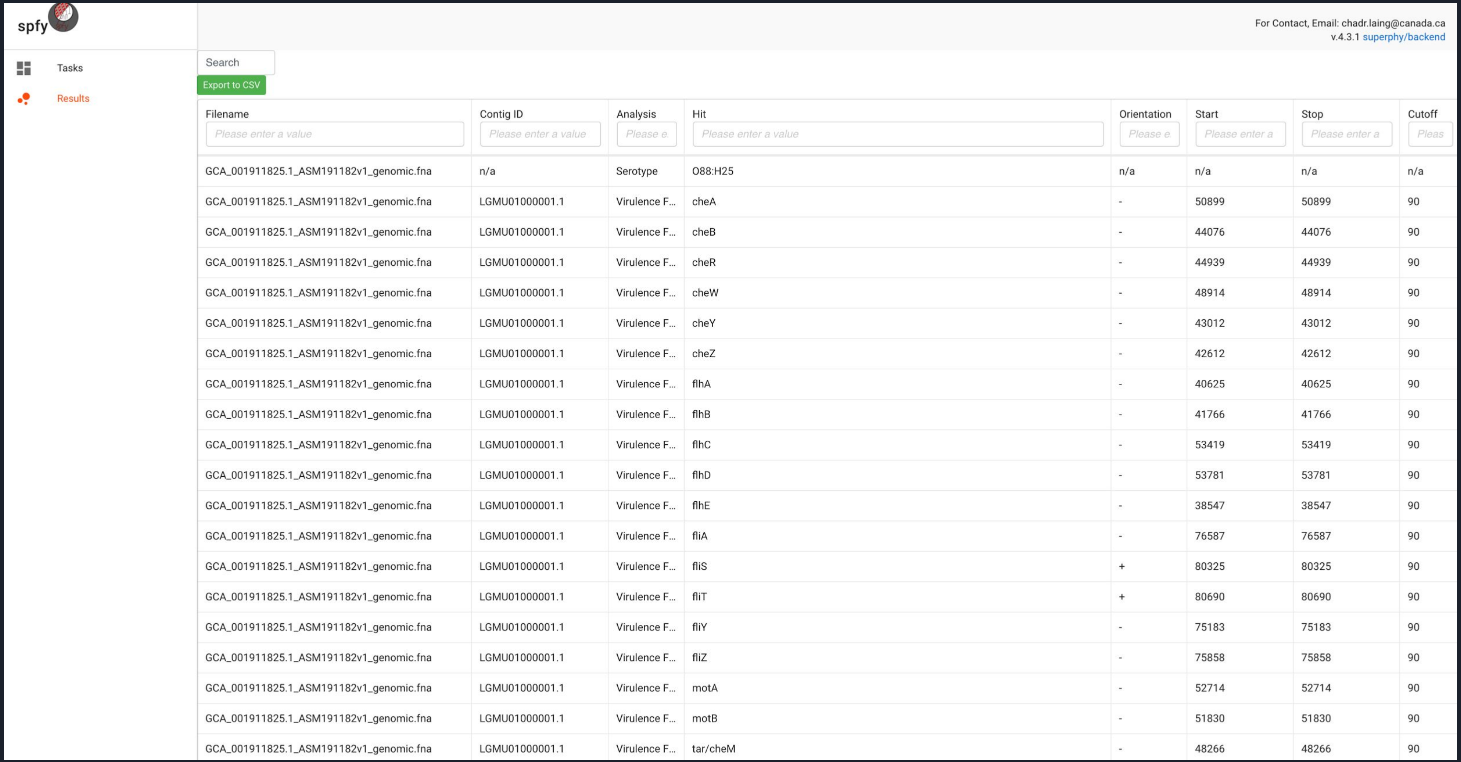
GCA_001891695.1_ASM189169v1_genomic.fna

GCA_001891995.1_ASM189199v1_genomic.fna

GCA_001901445.1_ASM190144v1_genomic.fna

Solution

Simplicity: one-click results

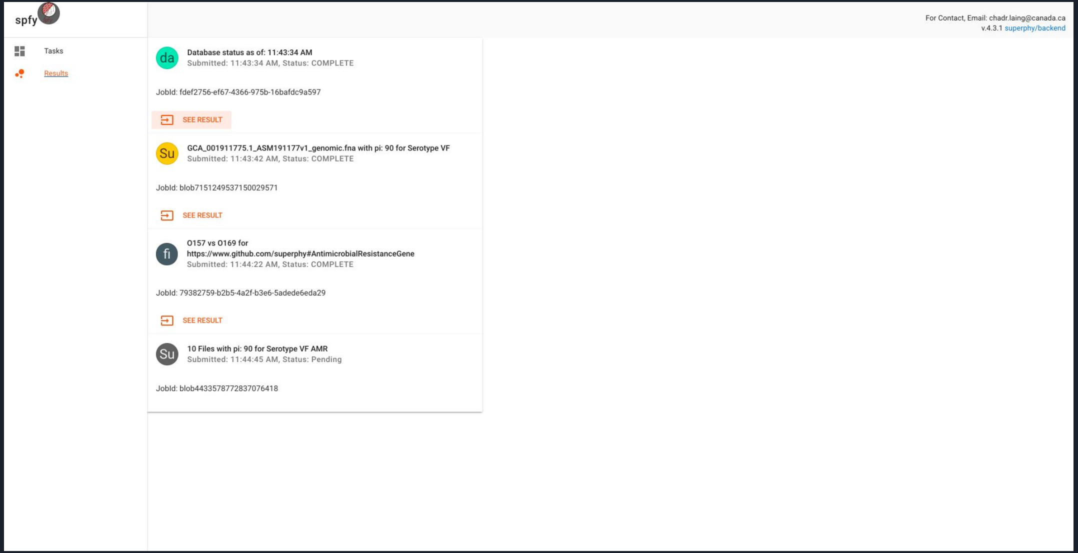


For Contact, Email: chadr.laing@canada.ca
v.4.3.1 superphy/backend

Filename	Contig ID	Analysis	Hit	Orientation	Start	Stop	Cutoff
Please enter a value	Please enter a value	Please e	Please enter a value	Please e	Please enter a	Please enter a	Pleas
GCA_001911825.1_ASM191182v1_genomic.fna	n/a	Serotype	O88:H25	n/a	n/a	n/a	n/a
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	cheA	-	50899	50899	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	cheB	-	44076	44076	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	cheR	-	44939	44939	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	cheW	-	48914	48914	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	cheY	-	43012	43012	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	cheZ	-	42612	42612	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	flhA	-	40625	40625	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	flhB	-	41766	41766	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	flhC	-	53419	53419	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	flhD	-	53781	53781	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	flhE	-	38547	38547	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	fliA	-	76587	76587	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	fliS	+	80325	80325	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	fliT	+	80690	80690	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	fliY	-	75183	75183	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	fliZ	-	75858	75858	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	motA	-	52714	52714	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	motB	-	51830	51830	90
GCA_001911825.1_ASM191182v1_genomic.fna	LGMU01000001.1	Virulence F...	tar/cheM	-	48266	48266	90

Solution

Simplicity: material design



For Contact, Email: chadr.laing@canada.ca
v.4.3.1 superphy/backend

da Database status as of: 11:43:34 AM
Submitted: 11:43:34 AM, Status: COMPLETE

JobId: fdef2756-ef67-4366-975b-16bafdc9a597

Su GCA_001911775.1_ASM191177v1_genomic.fna with pi: 90 for Serotype VF
Submitted: 11:43:42 AM, Status: COMPLETE

JobId: blob7151249537150029571

fi O157 vs O169 for
https://www.github.com/superphy#AntimicrobialResistanceGene
Submitted: 11:44:22 AM, Status: COMPLETE

JobId: 79382759-b2b5-4a2f-b3e6-5adede6eda29

Su 10 Files with pi: 90 for Serotype VF AMR
Submitted: 11:44:45 AM, Status: Pending

JobId: blob4433578772837076418

Card Based

Challenges deep-dive

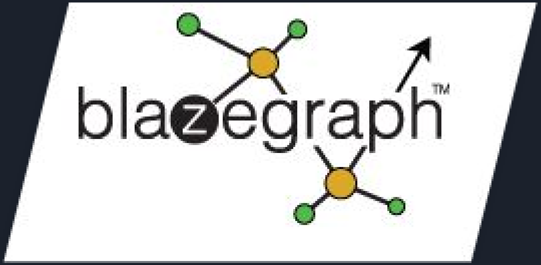
Goal: Scalability

- Quickly integrate new analyses & results
- Ability to perform group comparisons across thousands of genomes

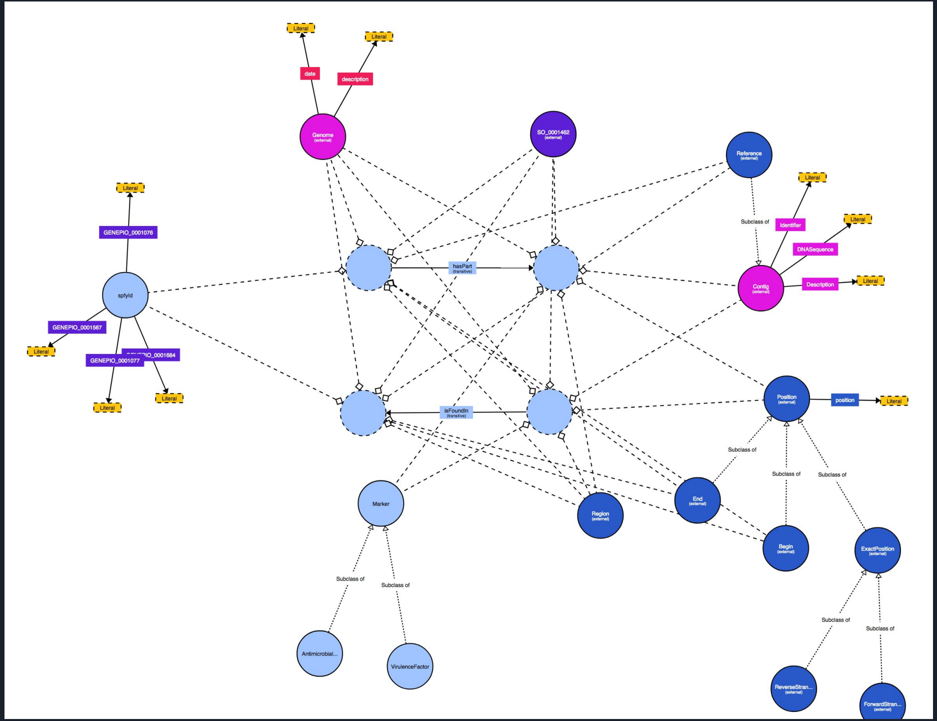
Solution: Technology

- Graph database
- Pandas, SciPy

Solution

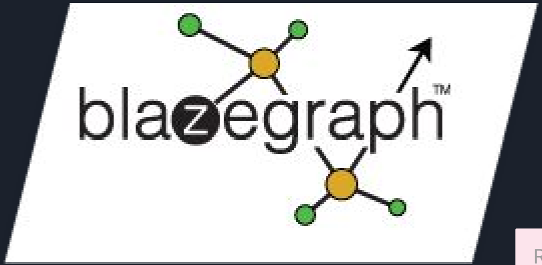


Scalability: graph database

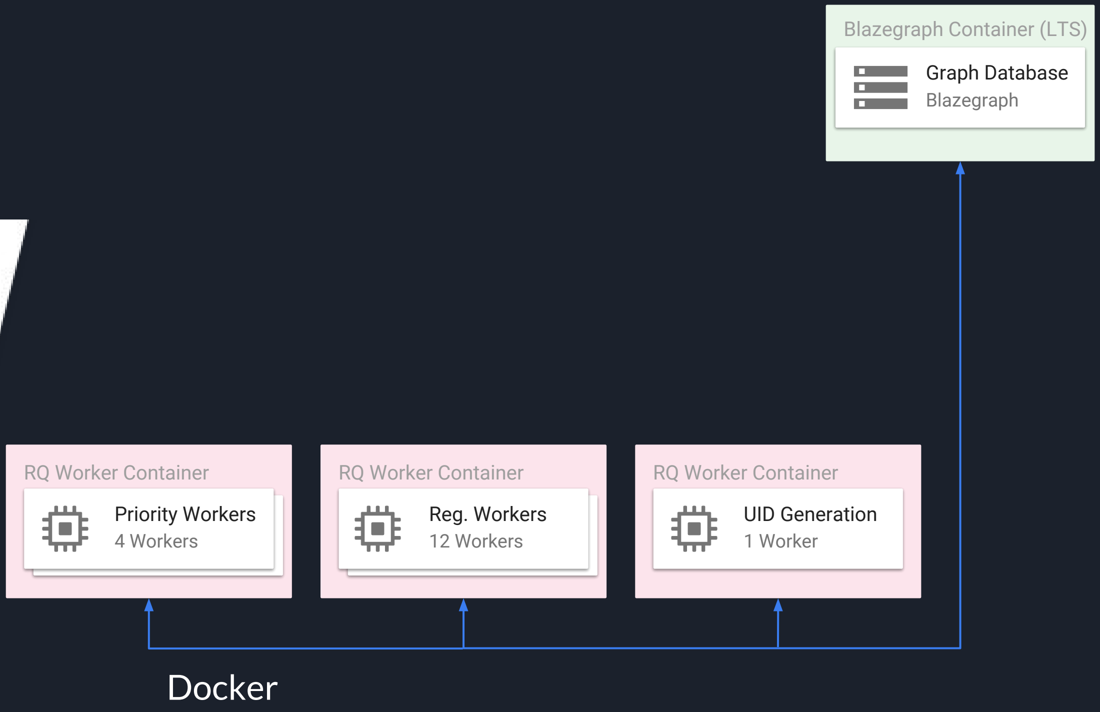


Semantic Web

Solution



Scalability: graph database



Blazegraph Container (LTS)

Graph Database
Blazegraph

RQ Worker Container

Priority Workers
4 Workers

RQ Worker Container

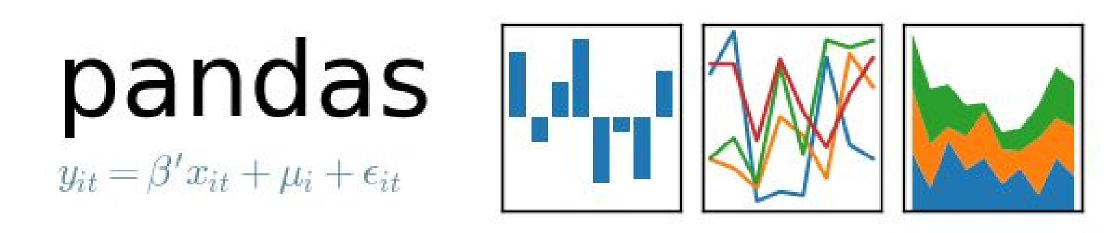
Reg. Workers
12 Workers

RQ Worker Container

UID Generation
1 Worker

Docker

Solution



Scalability: pandas, scipy

Python Wrappers to C Code

Challenges deep-dive

Speed

Modernize

- Task queues
- ReactJS

Simplicity

Familiar Design

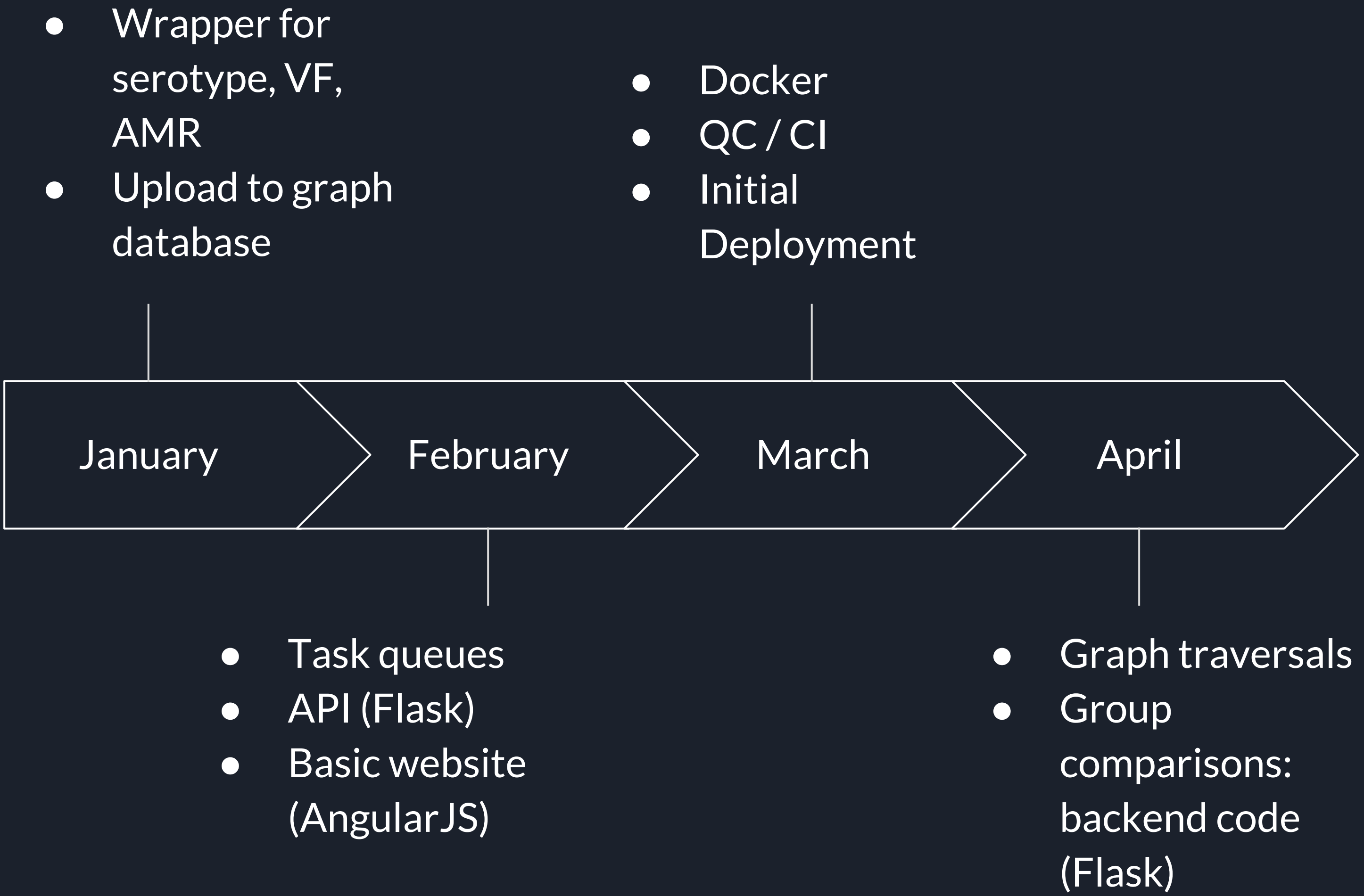
- One click results
- Material design (Google)

Scalability

Technology

- Graph database
- Pandas, SciPy

Timeline



January

- Wrapper for serotype, VF, AMR
- Upload to graph database

February

- Task queues
- API (Flask)
- Basic website (AngularJS)

March

- Docker
- QC / CI
- Initial Deployment

April

- Graph traversals
- Group comparisons: backend code (Flask)

May

- Group comparisons: frontend code (ReactJS)
- 2nd deployment

June

- Implemented subtyping in React
- Read the Docs

July

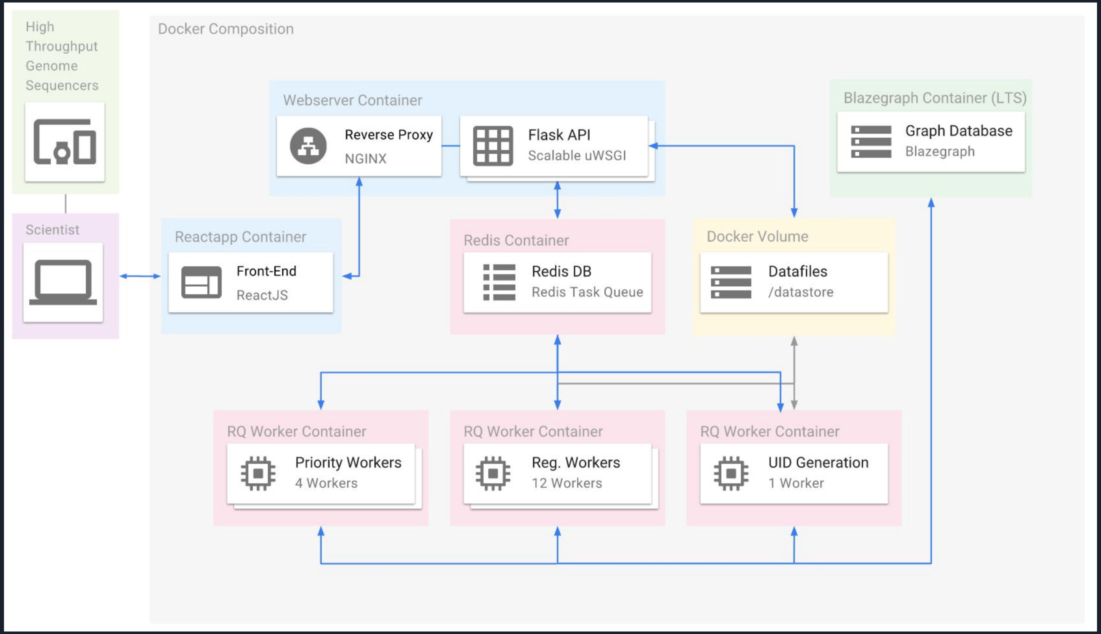
- Refactor backend
- Additional features
- 3rd deployment

August

- Paper writing

End Result

Platform Architecture



High Throughput Genome Sequencers

Scientist

Docker Composition

Webserver Container

Reverse Proxy
NGINX

Flask API
Scalable uWSGI

Blazegraph Container (LTS)

Graph Database
Blazegraph

Reactapp Container

Front-End
ReactJS

Redis Container

Redis DB
Redis Task Queue

Docker Volume

Datafiles
/datastore

RQ Worker Container

Priority Workers
4 Workers

RQ Worker Container

Reg. Workers
12 Workers

RQ Worker Container

UID Generation
1 Worker

Live Demo

End Result:

- Real-time serotype, VF, AMR prediction
- Within 2-3 minutes

- Storage and retrieval of genomes & results in a Graph Database
- Test set: 5353 GenBank genomes

- Live group comparisons of database entries within seconds
- Ex. O157 vs O53 for all known VFs