Building a data platform
For tomorrow’s health service

Natalie Godec,
Platform Engineer, Babylon Health
Networking
$ whoami 🦄

Natalie Godec
Babylon Health
DevOps (❤️ω❤️)
Platform Engineer
🔗 ouvessvit
Part I

Our mission
Babylon’s mission is to put an accessible and affordable health service in the hands of every person on Earth
Our go to market options

With governments

Value proposition extension

As employee benefit
Global
Part II

Data in AI-driven healthcare
What are we trying to achieve?

data

collect

store

analyse

insights

research
data driven decisions
Data Tech

kafka

nifi

hadoop
Data Tech

- Kafka
- NiFi
- Hadoop
- Google BigQuery
Enable teams to access, analyse and research on provided data, in an accessible, secure and auditable manner.
Part III
Infrastructure
Terraform!
Infrastructure components

- components
  - vpc-service-perimeter
  - gcp-aws-vpn
  - gcp-network
  - big-query
  - cloud-composer
  - gcp-buckets
  - gcp-projects
  - gcp-service-account
  - google-groups-access

- modules
  - gcp
    - big-query-data-transfer
    - big-query-dataset
    - big-query-table
    - bucket
    - project
    - service-account-bazaar
    - service-account-bq
  - aws
Infrastructure components

**components**
- vpc-service-perimeter
- gcp-aws-vpn
- gcp-network
- big-query
- cloud-composer
- gcp-buckets
- gcp-projects
- gcp-service-account
- google-groups-access

**modules**
- gcp
  - big-query-data-transfer
  - big-query-dataset
  - big-query-table
  - bucket
  - project
  - service-account-bazaar
  - service-account-bq
- aws
Infrastructure components

**components**
- vpc-service-perimeter
- gcp-aws-vpn
- gcp-network
- big-query
- cloud-composer
- gcp-buckets
- gcp-projects
- gcp-service-account
- google-groups-access

**modules**
- gcp
  - big-query-data-transfer
  - big-query-dataset
  - big-query-table
  - bucket
  - project
  - service-account-bazaar
  - service-account-bq
- aws
Infrastructure components

- **components**
  - vpc-service-perimeter
  - gcp-aws-vpn
  - gcp-network
  - big-query
  - cloud-composer
  - gcp-buckets
  - gcp-projects
  - gcp-service-account
  - google-groups-access

- **modules**
  - gcp
    - big-query-data-transfer
    - big-query-dataset
    - big-query-table
    - bucket
    - project
    - service-account-bazaar
    - service-account-bq
  - aws
So, we already have our pipeline for AWS. Can we just copy the setup?
Infrastructure Pipeline

Terraform → GitHub → build → deploy
AWS Infrastructure at Babylon

Accounts:

Nonprod  Prod  Internal
AWS Infrastructure at Babylon

Nonprod  Prod  Internal

eu-west-1
  VPC
  RDS
  S3
  EKS

can-central-1
  VPC
  RDS
  S3
  EKS

us-east-1
  VPC
  RDS
  S3
  EKS
>>> environment.tfvars

tenant         = "babylon"
region         = "eu-west-1"
env            = "prod"
location       = "uk"
vpc_cidr       = "10.4.5.6/16"
vpc_name       = "base-vpc"

>>> vpc.tf

module "base" {
    source = "../..//modules/aws/vpc/"
    name   = "${var.vpc_name}"
    cidr_block   = "${var.vpc_cidr}"
    team        = "platform"
    location    = "${var.location}"
    environment = "${var.env}"
    tenant      = "${var.tenant}"
}

*IP addresses are randomly generated*
AWS

account/
  ── location
  │   ├── environment1
  │   └── environment2
  └── environment3

nonprod/uk/preprod

GCP

account/
  ── location
  │   ── environment1
  │   │   └── project1
  │   │   └── project2
  │   └── project3
  └── environment2
    └── project1
    └── project2
GCP Access Model

Read DS1

DS1
ACL

DS2
ACL

project1
roles/bigquery.user
**AWS**

- account/
  - location
    - environment1
    - environment2
    - environment3

- nonprod/uk/preprod

**GCP**

- account/ team/
  - location
    - environment
      - project1
      - project2
      - project3
    - environment2
      - project1
      - project2

- team-project-env-location
Securing your APIs:
VPC Service Controls
vpc-service-perimeter/
  └── terraform
      ├── backend.tf
      │    ├── data.tf
      │    └── environments
      │         └── organization
      │                 ├── backend.tfvars
      │                 └── environment.tfvars
      ├── locals.tf
      ├── main.tf
      ├── outputs.tf
      ├── provider.tf
      └── vars.tf

VPC Service Perimeter

- BigQuery
  - IP1
  - IP2

- project1
- project2
- project3
A bit more terraform...
resource "google_access_context_manager_access_policy" "plc" {
  parent = "organizations/${var.organization_id}"
  title  = "Babylon GCP Service restriction policy"
}

resource "google_access_context_manager_service_perimeter" "service_perimeter" {
  parent = "accessPolicies/${google_access_context_manager_access_policy.plc.name}"  
  name   = "accessPolicies/012345/servicePerimeters/${var.perimeter_policy_name}"
  title  = "${var.perimeter_policy_name}"

  status {
    resources = ["${local.project_number_list}"]
    restricted_services = "${var.api_urls}" 
    access_levels = ["${google_access_context_manager_access_level.lvl.name}"]
  }
}
resource "google_access_context_manager_access_level" "lvl" {
  parent = "accessPolicies/${google_access_context_manager_access_policy.plc.name}"
  name   = "accessPolicies/012345/accessLevels/${var.perimeter_policy_name}"
  title  = "internal_vpc_and_jenkins"

  basic {
    combining_function = "OR"

    conditions {
      members = "${concat(var.svc_accs,var.users,var.groups)}"
    }

    conditions {
      ip_subnetworks = ["89.76.231.222/32", "0.140.64.220/32", "5.1.143.250/32"]*
    }
  }
}

*IP addresses are randomly generated using browserling.com/tools/random-ip
AWS to GCP through a perimeter?

AWS

VPN GW

account1

Microservices

VPN tunnel

GCP

VPN GW

project1

project2

project3

VPC Service Perimeter
AWS to GCP through a perimeter?

### AWS: route table

<table>
<thead>
<tr>
<th>Destination</th>
<th>Target</th>
<th>Status</th>
<th>Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.1.2.3/24</td>
<td>vgw-012345</td>
<td>active</td>
<td>Yes</td>
</tr>
<tr>
<td>10.4.5.6/16</td>
<td>vgw-012345</td>
<td>active</td>
<td>Yes</td>
</tr>
<tr>
<td>10.7.8.9/16</td>
<td>vgw-012345</td>
<td>active</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### GCP: route table

<table>
<thead>
<tr>
<th>Destination range</th>
<th>Next Hop</th>
<th>Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.2.0.1/16</td>
<td>141.224.170.66</td>
<td>prod-uk</td>
</tr>
<tr>
<td>10.20.1.1/16</td>
<td>141.224.170.66</td>
<td>Prod-uk</td>
</tr>
</tbody>
</table>

AWS: VPN Gateway
ID: vgw-012345

GCP: VPN Gateway
BGP Peer IP: 141.224.170.66

*IP addresses are randomly generated
Automate the automation

Terraform → GitHub → build → auto

deploy to dev

deploy to staging

GCP projects
Part IV

Data Engineering Workflow
Airflow, Composer and DAGs

Python operator → Branching → Python operator → SQL operator → Custom operator → end

Data_processing
- dags
  - UK_scripts
    - dataset1
    - dataset2
- plugins
  - control_plugin
    - operators
- schemas
  - dataset1
  - dataset2
Data Engineering Workflow

1. Analyse need
2. Map Data
3. Write DAGs
4. Deploy!

data analyst  data engineer
Fancy conclusion slide
Get in touch

linkedin.com/in/nataliia-godec/
instagram.com/ouvessvit/
twitter.com/Babylon_Eng
babylonhealth.com/blog/tech
Rate today’s session

Cyberconflict: A new era of war, sabotage, and fear

9:55AM - 10:10AM Wednesday, March 27, 2019
Location: Ballroom
Secondary Topics: Security and Privacy

Rate This Session

We’re living in a new era of constant sabotage, misinformation, and fear, in which everyone is a target, and you’re often the collateral damage in a growing conflict among states. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. Moving from the White House Situation Room to the dens of Chinese, Russian, North Korean, and Iranian hackers to the boardrooms of Silicon Valley, David reveals a world coming face-to-face with the perils of technological revolution—a conflict that the United States helped start when it began using cyberweapons against Iranian nuclear plants and North Korean missile launches. But now we find ourselves in a conflict we’re uncertain how to control, as our adversaries exploit vulnerabilities in our hyperconnected nation and we struggle to figure out how to deter these complex, short-of-war attacks.

David Sanger
The New York Times

David E. Sanger is the national security correspondent for The New York Times as well as a national security and political contributor for CNN and a frequent guest on CBS This Morning, Face the Nation, and many PBS shows.

Session page on conference website
O'Reilly Events App