Datasets Públicos de COVID-19 no Google Cloud Marketplace: Um guia abrangente para pesquisa de análise de políticas

Hierarquia de datasets e estrutura de arquivos

Os datasets de COVID-19 no Google Cloud estão organizados dentro do projeto `bigquery-public-data`, apresentando uma estrutura hierárquica sofisticada projetada para análise eficiente. O dataset principal **COVID-19 Open Data** contém mais de 15 tabelas especializadas com chaves geográficas consistentes que possibilitam integração perfeita.

Organização primária do dataset:

- **Projeto principal**: `bigquery-public-data` (armazenamento multi-região US)
- **Dataset central**: `covid19_open_data` com 15+ tabelas interconectadas
- **Hierarquia geográfica**: Sistema de quatro níveis (País → Estado/Província → Condado/Município → Localidade)
- **Sistema de chaves únicas**: Identificadores de localização padronizados resolvendo discrepâncias de códigos ISO/NUTS/FIPS em mais de 20.000 localizações globais

Os dados seguem formato de armazenamento colunar otimizado para análises BigQuery, com particionamento por data e clustering geográfico para performance. Cada tabela mantém campos `key` consistentes possibilitando junções diretas entre dados epidemiológicos, de políticas, mobilidade e econômicos.

Principais informações e conteúdo dos dados

O Google Cloud hospeda uma extensa coleção de tipos de dados COVID-19, tornando-se particularmente valioso para análise abrangente de políticas:

Dados epidemiológicos incluem séries temporais diárias de janeiro de 2020 em diante rastreando casos confirmados, mortes, recuperações, volumes de testagem, hospitalizações e taxas de vacinação em múltiplas granularidades geográficas. A plataforma integra dados da Johns Hopkins CSSE, ECDC, New York Times e ministérios de saúde de países individuais.

Dados de políticas e intervenções apresentam o Oxford Government Response Tracker (OxCGRT) completo com 19 indicadores de política padronizados medindo fechamentos de escolas, restrições no local de trabalho, proibições de viagem e medidas de apoio econômico. Cada política é pontuada em escalas consistentes com índices de rigor compostos (0-100) possibilitando comparações entre países.

Dados de mobilidade e comportamento dos Google Community Mobility Reports rastreiam mudanças de movimento em seis categorias de localização (varejo, mercearias, parques, transporte, locais de trabalho, residencial) comparadas a baselines pré-pandemia. Estes dados utilizam proteção de

privacidade diferencial e fornecem mudanças percentuais dos valores baseline de 3 de janeiro a 6 de fevereiro de 2020.

Datasets de apoio incluem indicadores econômicos, métricas de capacidade de saúde, perfis demográficos, condições climáticas e dados de tendências de busca - todos alinhados temporal e geograficamente para análise integrada.

Principais considerações e limitações para análise

Pesquisadores devem navegar várias limitações críticas ao usar estes datasets:

Restrições de atualização de dados: O Google descontinuou atualizações em tempo real em 15 de setembro de 2022, convertendo o repositório para retrospectivo apenas. O tracker de políticas de Oxford cessou coleta em maio de 2023, e relatórios de mobilidade terminaram em outubro de 2022. Pesquisadores analisando períodos recentes devem buscar dados suplementares.

Questões de qualidade e consistência variam significativamente por região. Desafios comuns incluem artefatos de relatórios em lote (onde países reportam mortes acumuladas em dias únicos), cobertura subnacional inconsistente, mudanças de convenções de nomenclatura no meio do dataset, e incompatibilidades de agregação onde totais regionais não somam aos níveis de país. Alguns países carecem de dados de testagem, coordenadas geográficas, ou cronogramas de relatórios consistentes.

Considerações metodológicas para análise de lockdown incluem contabilizar lags de 7-14 dias entre implementação de política e efeitos epidemiológicos, distinguir datas de anúncio versus aplicação de política, e controlar variações sazonais que confundem padrões de mobilidade. O período baseline de mobilidade pode não representar padrões normais para todas as regiões, particularmente aquelas com restrições pré-existentes significativas.

Restrições legais e éticas limitam uso para fins educacionais e de pesquisa apenas, proibindo explicitamente aplicações médicas. Todos os datasets requerem atribuição adequada sob licenciamento CC BY 4.0, com requisitos adicionais de Termos de Serviço do Google para dados de mobilidade e busca.

Especificações técnicas

A plataforma aproveita a infraestrutura avançada de analytics do BigQuery com parâmetros técnicos específicos:

Métodos de acesso incluem consultas SQL através do console BigQuery (com 1TB de nível gratuito mensal), API REST com bibliotecas cliente para linguagens principais de programação, downloads diretos CSV/JSON do Cloud Storage, e integração limitada do Google Sheets para datasets menores.

Formatos de dados e esquema utilizam armazenamento colunar nativo do BigQuery (formato Capacitor) com compressão automática e replicação multi-zona. Tabelas apresentam esquemas fortemente tipados com tratamento NULL apropriado, suportando atualizações incrementais através de pipelines ETL sofisticados documentados no repositório GitHub open-source.

Otimização de performance beneficia-se de particionamento inteligente por colunas de data e clustering por chaves geográficas, possibilitando consultas eficientes em bilhões de linhas. Processamento de consulta paraleliza automaticamente na infraestrutura do Google, com cache de resultados reduzindo custos para análises repetidas.

Exemplo de consulta para análise de efetividade de lockdown:

```sql
WITH policy_timeline AS (
  SELECT
    country_name,
    date,
    cumulative_confirmed,
    stay_at_home_requirements,
    LAG(cumulative_confirmed, 14) OVER (
      PARTITION BY country_name ORDER BY date
    ) AS cases_two_weeks_prior
  FROM `bigquery-public-data.covid19_open_data.covid19_open_data`
  WHERE date BETWEEN '2020-03-01' AND '2020-06-01'
)
SELECT
  country_name,
  date,
  stay_at_home_requirements,
  ((cumulative_confirmed - cases_two_weeks_prior) /
    NULLIF(cases_two_weeks_prior, 0)) * 100 AS two_week_growth_rate
FROM policy_timeline
WHERE cases_two_weeks_prior > 100
```

Integração com outros datasets de COVID-19

A plataforma se destaca na integração de múltiplas fontes autoritativas, embora pesquisadores devam entender as relações:

Integrações diretas incluem o tracker de Oxford incorporado nas tabelas Open Data, dados da Johns Hopkins disponíveis como tabelas separadas do BigQuery, e dados próprios de mobilidade e tendências de busca do Google. Estes compartilham chaves geográficas consistentes possibilitando junções diretas.

Datasets externos complementares como Our World in Data requerem importação manual mas frequentemente fornecem pontos de validação ou preenchem lacunas geográficas. Pesquisadores comumente combinam dados do Google Cloud com outputs de modelagem universitária, escritórios de estatísticas nacionais, ou datasets especializados de saúde.

Melhores práticas de integração envolvem usar a tabela mestre `index` para padronização geográfica, alavancando formatos de data consistentes em todas as tabelas, e validando totais em diferentes níveis

de agregação. O código de pipeline open-source fornece transparência nos processos de transformação de dados.

Recomendações práticas para pesquisadores de políticas

Para pesquisadores analisando timing e efetividade de lockdowns, recomendo começar com três tabelas centrais: `oxford-government-response` para medidas políticas, `epidemiology` para trajetórias de casos, e `mobility` para verificação comportamental. Focalize análises iniciais em países com relatórios consistentes e documentação completa de políticas.

Fatores críticos de sucesso incluem contabilizar lags de implementação em seus modelos, usar médias móveis para suavizar flutuações diárias, comparar regiões similares com timings de política diferentes como experimentos naturais, e sempre validar achados contra fontes originais de dados.

Estratégias de otimização de custos: Estruturar consultas para aproveitar particionamento por data, selecionar apenas colunas requeridas para minimizar processamento de dados, usar cláusulas WHERE cedo para reduzir volumes de escaneamento, e utilizar o nível gratuito do BigQuery efetivamente através de batching de consultas.

A combinação de cobertura abrangente, indicadores de política padronizados e dados de mobilidade integrados torna esta plataforma unicamente valiosa para pesquisa de políticas de COVID-19, apesar de requerer atenção cuidadosa a questões de qualidade de dados e limitações temporais. Pesquisadores que investem tempo entendendo a estrutura e limitações destes datasets os acharão inestimáveis para análise rigorosa de efetividade de políticas.

Resumo das principais tabelas disponíveis

Tabelas principais do dataset `covid19_open_data`:

1. **`covid19_open_data`** - Tabela principal com séries temporais diárias
   - Casos confirmados, mortes, recuperações
   - Dados de vacinação e testagem
   - Indicadores de políticas (Oxford tracker)
   - Chave geográfica padronizada
2. **`epidemiology`** - Dados epidemiológicos puros
   - Casos novos e cumulativos
   - Mortes novas e cumulativas
   - Dados de testagem quando disponíveis
3. **`oxford_government_response`** - Medidas governamentais
   - 19 indicadores de política
   - Índices de rigor (0-100)

- Datas de implementação

4. `mobility` - Relatórios de mobilidade Google
   - 6 categorias de localização
   - Mudanças percentuais vs baseline
   - Proteção de privacidade diferencial
5. `demographics` - Dados demográficos
   - População, densidade
   - Distribuição etária
   - Indicadores socioeconômicos
6. `economy` - Indicadores econômicos
   - PIB per capita
   - Índices de desenvolvimento
   - Gastos em saúde
7. `health` - Capacidade do sistema de saúde
   - Leitos hospitalares per capita
   - Profissionais de saúde
   - Gastos com saúde
8. `weather` - Dados meteorológicos
   - Temperatura, umidade
   - Precipitação
   - Qualidade do ar

Pontos de atenção específicos:

⚠️ **Atualizações descontinuadas**: Dados param em setembro/outubro 2022 ⚠️ **Qualidade variável**: Validar sempre dados por região/país ⚠️ **Lags temporais**: Políticas vs efeitos epidemiológicos (7-14 dias) ⚠️ **Baseline de mobilidade**: Pode não ser representativa para algumas regiões ⚠️ **Agregações**: Totais regionais nem sempre somam aos nacionais