Lepton AI All Features

- Observability
  - Monitor
    - Create
      - name
      - alert
      - url
      - body
      - header
      - expected status
      - expected regex match
      - check interval
      - max latency
      - consecutive success count
      - consecutive failure count
    - Monitoring
      - Mute
      - Unmute
      - Delete
      - Edit
    - Alert
      - Name
      - webhook
  - Logs
    - Scope
    - Replicas
    - Search
    - Load older
    - Load newer
- Settings
  - Plans
    - Basic
    - Standard
    - Enterprise
  - Members
    - Add Member
      - Role
      - Email
    - Delete
    - Edit role
  - Tokens
    - User token
    - Workspace token
  - Secrets
  - Registries
  - Audit Logs
- Billing
  - Usage
    - API
    - Storage
    - Compute
  - Invoice
    - open invoice
    - upcoming invoice
    - history invoices
  - Pay
    - Postpaid
      - payment method
      - credits
    - Prepaid
      - credits
      - auto payment
      - payment method
- Machines
  - Node Group
    - Total nodes
    - Total CPU
    - Total GPU
    - Total Memory
  - Metric
    - Public Network I/O
    - Private Network I/O
  - Nodes
    - List
      - Name
      - Status
      - CPU
      - GPU
      - Memory
      - System
    - Detail
      - Overview
      - Metrics
- Network
  - Ingress
    - Create
      - domain
      - region
      - admin
        - lepton.model.api
        - remove authorization header
    - Manage
      - Endpoint
        - deployment
        - weight
        - load balance policy
        - least request
      - Shadow endpoint
        - deployment
        - percent of traffic
    - Delete
- Storage
  - File System
    - Upload File
      - Rsync
      - CLI
      - Web
      - Cloud
    - New Folder
    - Delete
    - Copy Path
  - Object Storage
    - Public
    - Private
  - KV
    - new namespace
    - delete namespace
  - Queue
    - new
    - delete
- Compute
  - Deployment
    - Create
      - photon
        - Photon
          - pre-built
            - lsf
            - vllm
            - public photons (expires)
          - custom
            - dev
              - write
                - dependencies
                - extra files
                - docker images
              - debug
              - package and deploy
            - manage
              - local
              - remote
      - name
      - photon
      - image
        - Container image
        - Private image registry auth
        - Start command
      - resource
        - replicas
        - resource shape
        - node group
        - auto scale
          - scale replica to zero when no traffic more than
          - based on traffic
            - QPM per replica
            - path filter
            - method filter
            - min replica
            - max replica
          - based on GPU utilization
            - threshold
            - min replica
            - max replica
      - access control
        - access tokens
          - deployment token
          - workspace token
          - public access
      - environment variables
        - env name
        - env value
          - secret
          - text
        - include workspace token
      - file system mount
        - file system
        - mount from
        - mount as
        - cache size
      - advanced
        - share memory
        - initial delay seconds
        - require approval to make replicas ready
        - disable pulling metrics from replica
        - header-based replica routing
      - info
        - visibility
    - Operate
      - Demo
        - Schema
        - Chat
      - API Docs
      - Replicas
        - Id
        - Machine
        - Created at
        - Last Termination
        - Status
        - Actions
          - Terminal
          - Logs
          - Metrics
            - QPS
            - Latency
            - CPU & Memory Util
            - Memory
            - GPU Memory
            - GPU Temperature
            - Network I/O
            - Disk I/O
            - File Descriptors
          - Delete
            - risk tips
      - Metrics
        - QPS
        - Latency
        - GPU Memory
        - GPU Temperature
        - HTTP Request Status
      - Logs
      - Timeline
    - Manage
      - Stop
      - Start
        - replica
        - auto scale
      - Restart
      - Update
        - Editable
          - photon version
          - replica
          - resource shape
            - admin only
          - access tokens
          - auto scale
          - environment variables
          - visibility
          - require approval to make replicas ready
          - disable pulling metrics from replica
          - header-based replica routing
        - Rolling update behavior
      - Delete
      - Clone
  - Job
    - Create
      - Best practice
      - Templates
      - Features
        - name
        - run command
        - container ports
        - visibility
        - share memory
        - max replica failure retry
        - max job failure retry
        - resource
          - replicas
          - resource shape
          - node group
        - image
          - Container image
          - private image registry auth
        - environment variables
          - env name
          - env value
            - secret
            - text
          - include workspace token
        - file system mount
          - file system
          - mount from
          - mount as
          - cache size
    - Manage
      - Clone
      - CLI Command
      - Delete
    - Operate
      - Replicas
        - ID
        - DNS Name
        - Created At
        - Last Termination
        - Status
        - Actions
          - Terminal
          - Logs
        - Metrics
          - QPS
          - Latency
          - CPU & Memory Util
          - Memory
          - GPU Memory
          - GPU Temperature
          - Network I/O
          - Disk I/O
          - File Descriptors
  - Pod
    - Create
      - GPU Type
      - name
      - resource
        - replicas
        - resource shape
          - gpu count
          - gpu type
        - node group
      - image
        - Container image
        - private image registry auth
        - ssh public key
        - jupyterlab
      - environment variables
        - env name
        - env value
          - secret
          - text
        - include workspace token
      - file system mount
        - file system
        - mount from
        - mount as
        - cache size
      - visibility
    - Manage
      - Stop
      - Restart
      - Delete
    - Operate
      - Connect
      - Terminal
      - Logs
      - Metrics
        - QPS
        - Latency
        - CPU & Memory Util
        - Memory
        - GPU Memory
        - GPU Temperature
        - Network I/O
        - Disk I/O
        - File Descriptors
  - Tune
    - Create
      - model
      - name
      - early stop loss
      - epochs
      - training data
        - upload
        - storage
    - Manage
      - Deploy
      - Stop
    - Operate
      - Try it out

Presented with xmind