OWLNEST® Spydernest

Proactive discovery and insights from contexts

OWLNEST Corp.
주소 : 서울특별시 강남구 자곡로 174-10 강남에이스타워 G9 318호 (06373)
문의 : contact@owl-nest.com, +82 2-742-3021

OWLNEST® Spydernest System Requirements

To learn more about OWLNEST® Spydernest, please visit http://owl-nest.com/technology, test the applications and see the references.

집약된 웹데이터 수집 기술을 탑재한 데이터 크롤링 솔루션으로써, 비정형 웹문서에 존재하는 메타정보와 내용 정보를 빠르게 수집할 수 있습니다.

- 사이트 구조와 내용을 탐지하고, 추출 대상 항목별로 정확하게 데이터를 추출
- 구조 변화, 사이트 추가 등의 이슈 대응을 위한 룰셋 패키지 구조 제공

What?

집약된 데이터 수집 기술
Spydernest

Benefits

1	2	3
집약된 자연어처리 기술과 알고리즘	*문서 내용, 첨부파일, 링크 통합 처리*	*데이터 전체 집합의 구조를 간략히 확보*
개발 시간의 절약	***수집 사이트의 정보 통합 제공***	***비즈니스 기회로 집중***

Details

수집 기술 집약형 SW	수집 정보의 통합	비즈니스적 가치
• *키워드 문법표지 부착, 개체명 인식, 구문단위 추출, 자동 띄어쓰기, 문장단위 분할*	• *개별 수집 대상 정보에 대한 기능단위 분할처리기능 제공*	• *분석 또는 서비스를 위한 수집 데이터를 안정적으로 제공*
• *의존관계 추출 등의 고수준 기능을 자동화하여 효과적으로 제공*	• *수집 사이트 정보 통합 제공으로 서비스에서 필요로 하는 기능을 편리하게 이용*	• *데이터 가치의 향상을 위해 비정형 데이터를 반정형 데이터 테이블 형태로 저장*
• *사용자는 분석업무에 집중, 가치발견의 가능성 향상*	• *웹스케일 수준의 관련 데이터에 실시간으로 접근하여 정확한 항목을 수집*	• *비정형 텍스트 데이터를 이용하여 서비스 및 요구사항에 연계*

OWLNEST® Spydernest를 도입함으로써 집약된 웹데이터 수집 기술을 탑재한 데이터 크롤링 솔루션의 표준 프레임워크 기반 시스템 개발이 가능합니다.

- 사이트 구조와 내용을 탐지하고, 추출 대상 항목별로 정확하게 데이터를 추출
- 구조 변화, 사이트 추가 등의 이슈 대응을 위한 룰셋 패키지 구조 제공

주요 영역	핵심 요구사항
시스템 구축	**Spydernest 도입 기반 시스템 구현** • 집약된 수집 기술 노하우의 패키지 이용 • 데이터 수집 자체에 개발 자원 집중 • 인증, 수집, 적재, 조회, 성능 등 비용 최적화 • 숙련된 인력을 통한 개발 기간 최소화 • 기술 교육, 장애 대응 지원 • 신규 수집 개발시 Time-To-Market 확보
시스템 확장	
시스템 관리	**외부 오픈소스 기반 구현** 오픈소스 공개 기술 활용 개발 • 모든 요소 모듈의 직접 구현 필요 (인증 등) • 데이터 수집 이외 단순 기능 요소에 개발 자원 소요 • 유지보수 문제, 프레임워크 개발 자원 소요

▶ 핵심성공 요소

‘표준 + 확장’ 프레임워크

1. 표준 개발 패키지 제공
2. 브라우저 관리 표준 기능
3. 신규 수집기 구축시 확장성

데이터 중심의 개발 편의성 제공

4. 완성형 수집기 기본 기능
5. 데이터 의미 중심 편의성 제공
6. 개발자를 위한 인터페이스

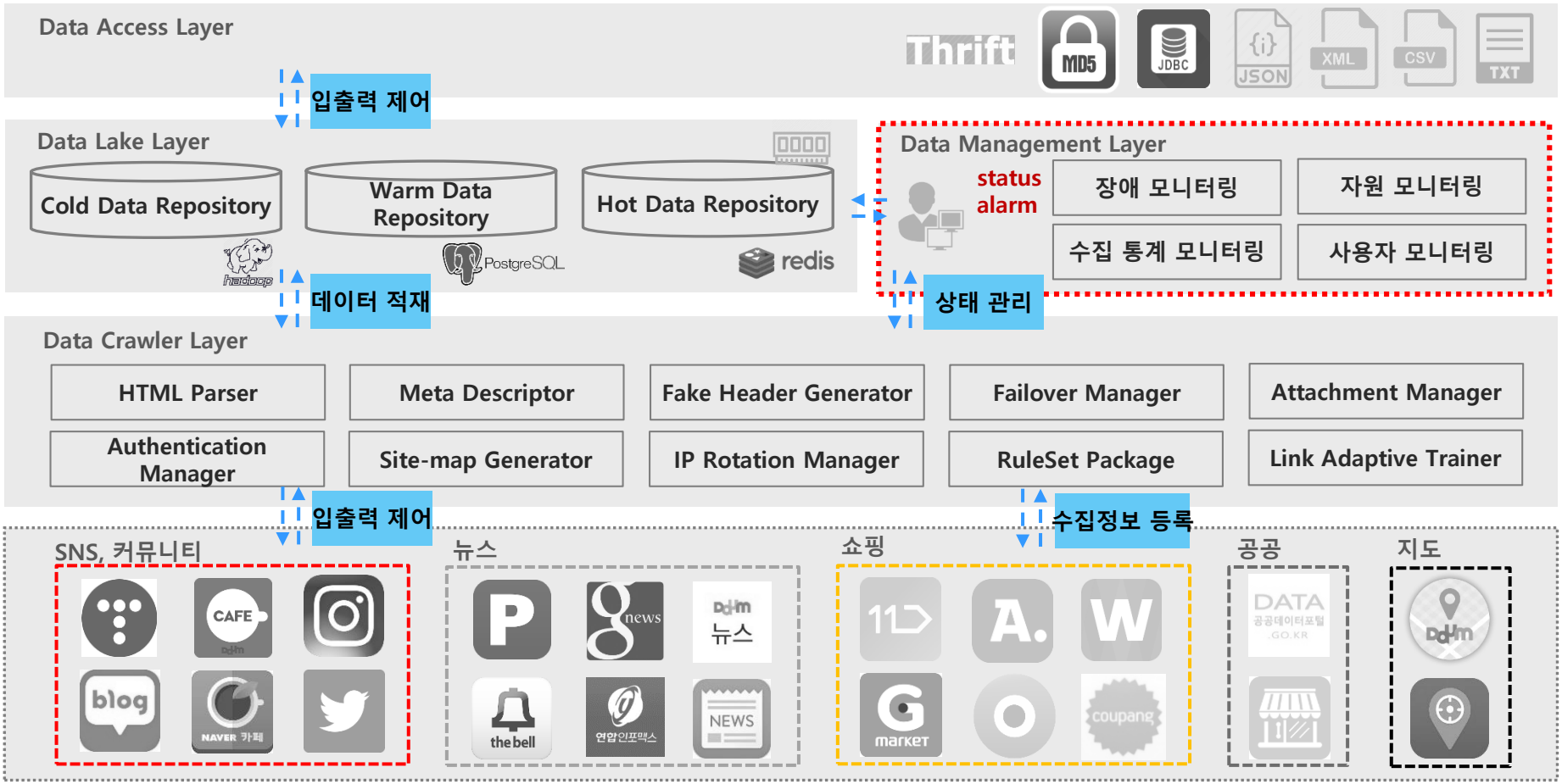
수집기 관리 기능

7. 수집 내역 조회 기능
8. 수집기 제어 기능

OWLNEST NLP Solutions

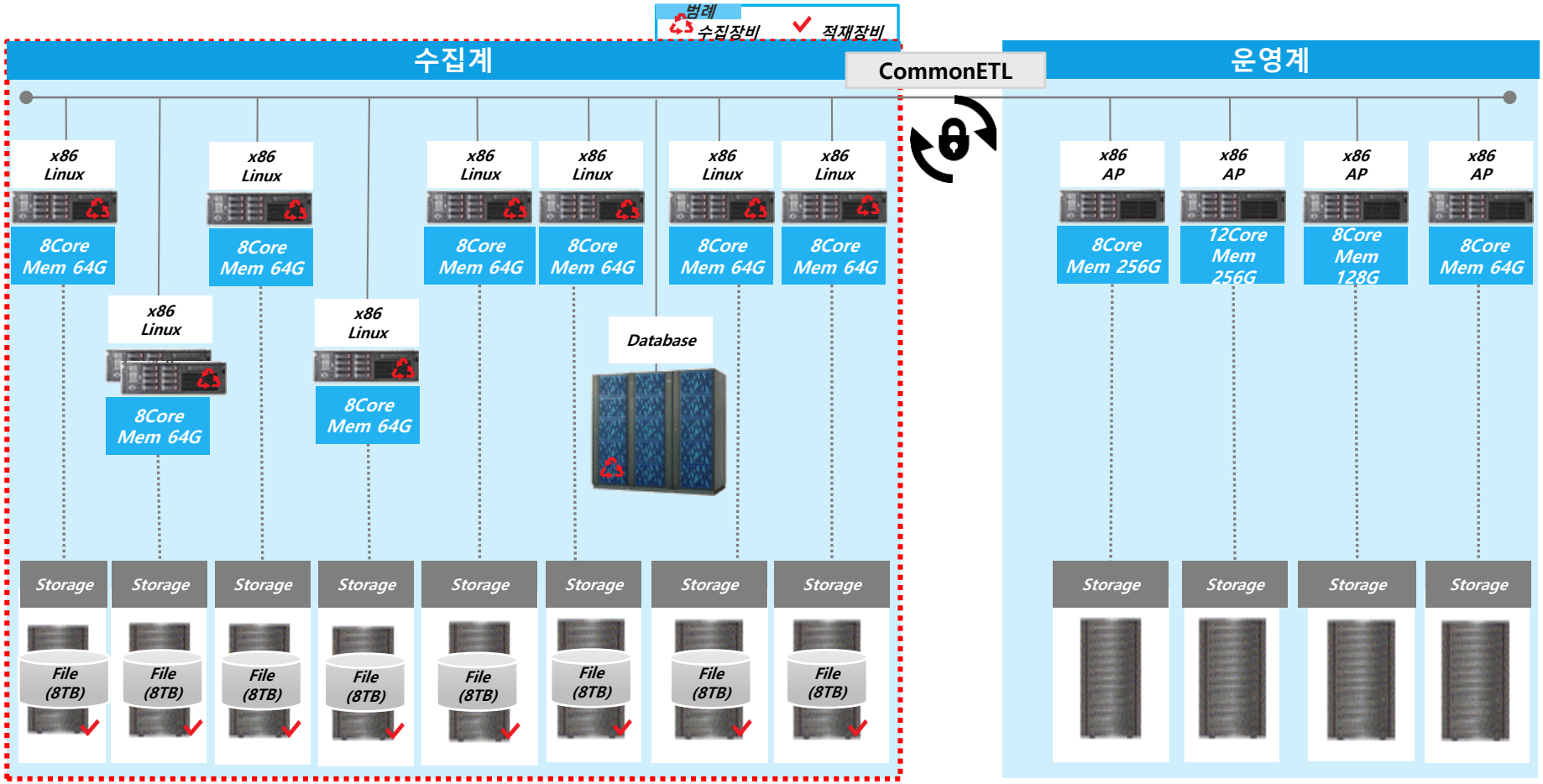
웹데이터 자동 수집을 위한 데이터 자동 수집 솔루션 Spydernest 의 기능을 기반으로 견고한 크롤링 아키텍처를 제공합니다.

- SNS, 커뮤니티, 뉴스, 쇼핑, 공공데이터, 지정데이터 등의 외부 비정형 데이터를 안정적으로 공급할 수 있는 체계적인 데이터 수집 기능을 제공



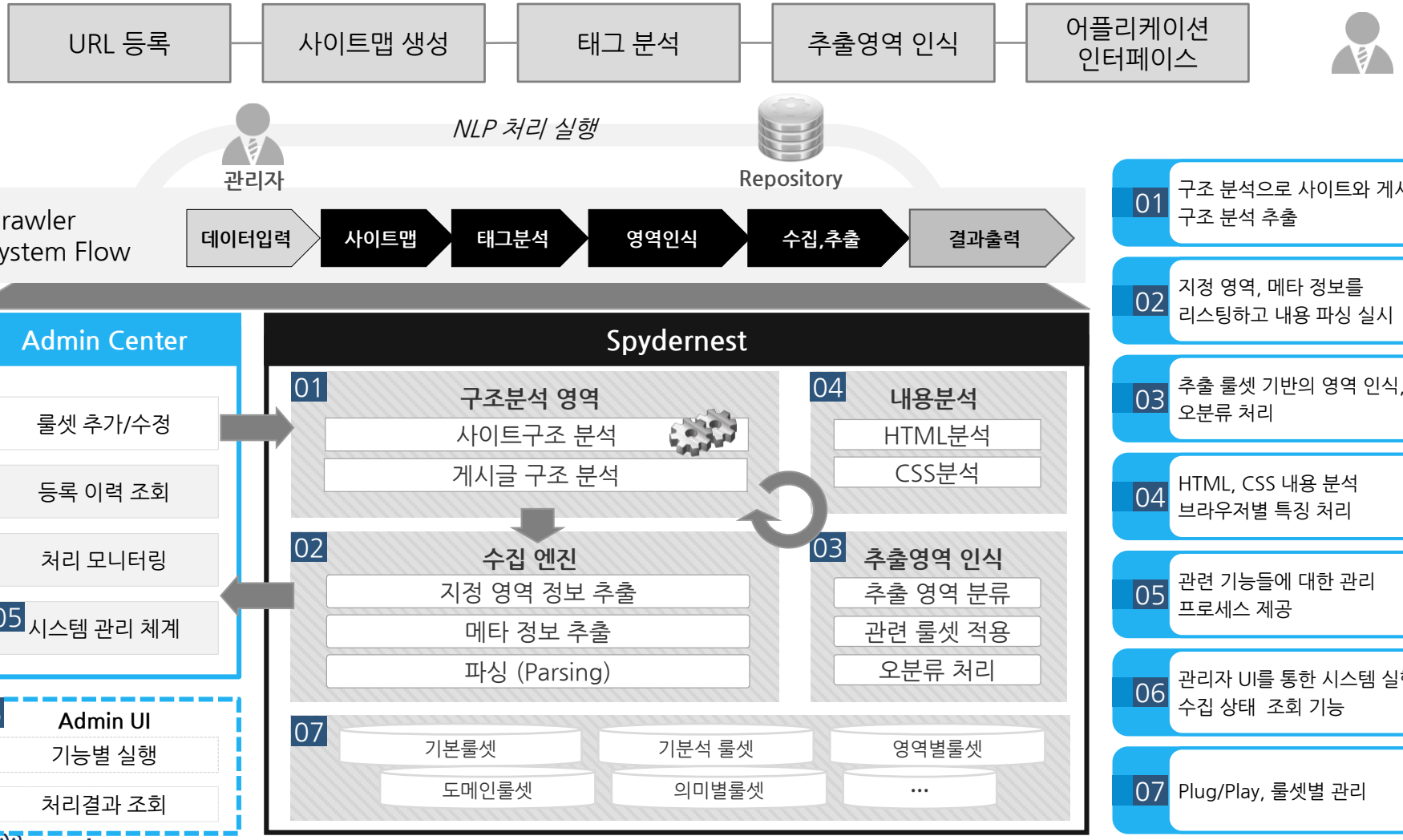
웹데이터 수집기의 안정적인 기능 구동을 위한 수집계 장비, 데이터베이스 장비의 운영 체계를 구성할 수 있도록 인터페이스 프로그램 CommonETL 을 함께 제공

- 수집기의 서비스 연결시 발생하는 일반 인터페이스 기능을 기본 지원
- 기능 : ETL, JDBC, Parallel Controller, Connection Pool, Encryption, Cross-Language



Spydernest 의 주요 구성요소와 처리 흐름은 아래와 같습니다.

- HTML Parser, Meta Descriptor, Fake Header Generator, Attachment Manager, Athentication Manager, Site-map Generator, IP Rotation Manager, RuleSet Package, Link Adaptive Trainer



01 구조 분석으로 사이트와 게시글 구조 분석 추출

02 지정 영역, 메타 정보를 리스팅하고 내용 파싱 실시

03 추출 룰셋 기반의 영역 인식, 오분류 처리

04 HTML, CSS 내용 분석 브라우저별 특징 처리

05 관련 기능들에 대한 관리 프로세스 제공

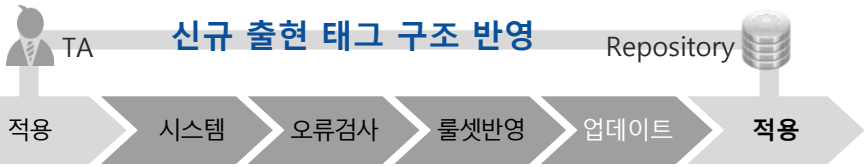
06 관리자 UI를 통한 시스템 실행 및 수집 상태 조회 기능

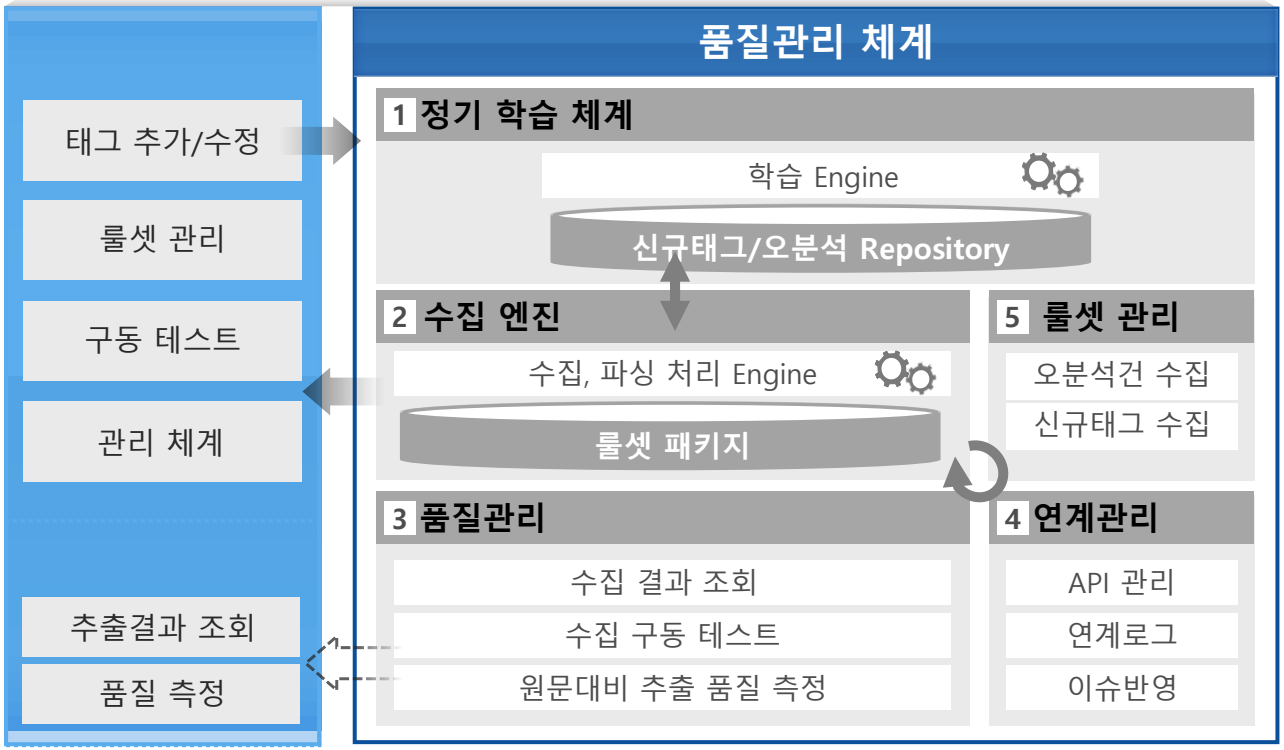
07 Plug/Play, 룰셋별 관리

Spydernest 솔루션 도입 후 품질관리 체계는 아래와 같습니다.

- 데이터 수집 솔루션 도입 후 자동 수집 품질을 효과적으로 유지하기 위해, 기술지원을 실시하여 신규 출현 태그를 반영하여 룰셋 패키지을 강화하고 수집 품질을 향상

데이터 수집 솔루션 품질운영 프로세스

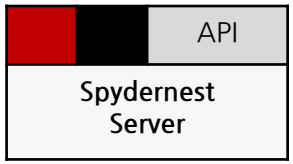


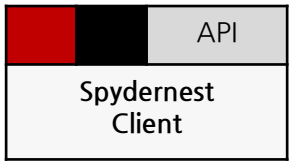


1. - 웹데이터 신규 출현 태그의 정기적 수집 Engine 반영
2. - 룰셋 기반 HTML, CSS, script 구문 일원화 처리
3. - 룰셋 업데이트 후 기준데이터 분석 품질 테스트로 품질관리
   - 추출 품질 측정
4. - API와 연계로그 관리
   - 실시간 이슈 반영
5. - 엔진 업데이트를 위한 오분석건, 신규태그 등을 수집

Spydernest 의 주요 특징은 아래와 같습니다.

- *시스템을 위한 서버 설치형 소프트웨어*
- *서버의 하드웨어 성능을 극대화할 수 있는 기업형 솔루션*





- *개발자를 위한 클라이언트 API*
- *각 컴포넌트 모듈에 대한 외부 제어 가능형 사용자 솔루션*

통합 인증 처리

• SSO 등의 로그인, 인증 보안접속 후 정보 수집을 위한 통합 인증 처리 기능 지원

웹페이지 로그인 정보를 저장한 브라우저 운용으로 통합 인증 유지 기능

IE, Chrome, Firefox 등의 특정 브라우저 종속 페이지 자동 접근 제어 기능

자동 인증, 접속

룰셋 패키지

• 수집 대상 페이지별 태그 정보를 수집하여 룰셋 패키지로 컴포넌트화 제공

컴포넌트 통합형 룰셋 제어구조로 패키지 관리의 편리성

개발자 친화적인 Key:Value 테이블 형태의 자료 관리 구조

개발 로직 통합

메타정보 수집

• 웹페이지 정보를 자동 분석하여 메타정보를 추출하고 저장

중요 메타정보의 일반/특수/확장 정보 자동 수집

개체-표현 간의 의존관계의 자동 추출을 통한 메타정보화

구조적 정보

접속정보 관리

• IP, Header 정보를 자동으로 우회하고 생성하는 정보 관리 기능 지원

Humanoid 스타일로 접속 Header 정보를 생성하는 차단 방지 기능

VPN 제공시 IP Rotation 기능

접속정보 관리

첨부파일 수집

• 웹페이지상의 파일, 이미지, 동영상 등을 수집, 텍스트 추출, 저장 기능 지원

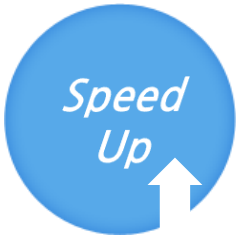
첨부파일 링크 다운로드 기능

문서 내 텍스트 추출 기능

고속 병렬 처리

Spydernest 은 기능유형으로 구분된 각 룰셋 패키지를 모듈구조로 관리합니다.

모듈형 룰셋 패키지 구조의 효과



도입 시스템의 안정화 주기 단축
빠른 시스템 안정화

손쉬운 룰셋 관리 및 유지보수
유지보수 효율성 향상

요구사항에 따라 룰셋 재조합
패키지 생성 효율화

일원화 구조 모듈의 위험요소
오류발생 Risk 최소화

기능유형	룰셋 유형								최소개수
Title	제목구조1	제목구조2	…	폰트정보					3
Content	내용구조1	내용구조2	…	폰트정보	테이블	블록정보	댓글정보	인코딩	7
Reply	댓글구조1	댓글구조2	…	작성일시	추천수	작성자	연결구조		6
Meta	페이징	작성일시	조회수	추천수	게시판	작성자	아바타	댓글정보	8
Link	참조링크	이미지	동영상	오디오	첨부파일	기타			5
총계									29

Spydernest 의 시스템 도입 시, 손쉬운 시스템 운영 및 유지보수를 위한 관리 프로세스를 제공합니다.

URL 입력 → Site Map 생성 → 태그 구조 분석 → 주요 태그 인식 → 어플리케이션 인터페이스

신규 룰셋 등록

추가태그 확보 → 추가태그 등록

- 신규 태그 분석을 통해 추가 대상 선정
  - 신규 추가태그 대상 선정요소
  - 신규 추가태그 후보 선정
  - 신구 츄가태그 적용 전략 수립
- 신규 추가태그 테이블 작성

데이터 관리 / 컴파일

데이터 추가/수정 → 룰셋 생성

- 데이터 추가/수정
- 룰셋 등록형태 오류 검증

시스템 운영

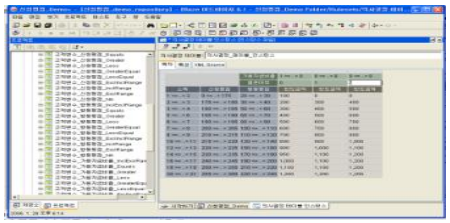
상태 저장 시스템 적용 → 운영/모니터링

변경 필요 시

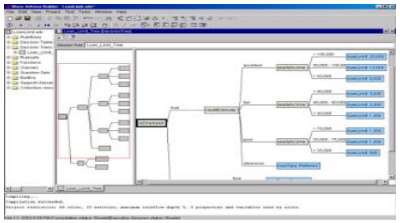
- 신규 등록 룰 포함 컴파일 수행
- 신규 룰셋 생성 후 시스템 적용

관리 완료 시 → 룰셋 패키지 반영

추가 태그 확보



룰셋 등록 테이블



테이블 구조 조회

등록 완료

프로젝트의 성공적인 목표 달성을 위한 Spydernest 활용 사례입니다.

프로젝트 명	발주처	기간	역할	형태
에뛰드하우스 Voice Mining Program 구축	아모레퍼시픽	'19.01. ~ 현재	시스템 구축	서비스 제공
신한금융그룹 그룹공통 빅데이터플랫폼 구축	신한금융그룹	'18.06. ~ '18.12.	시스템 구축	솔루션 납품
텍스트 데이터 내 재난안전 정보탐색 알고리즘 개선 및 탐색영역 확대 기술개발	국립재난안전연구원	'18.06. ~ '18.12.	시스템 개발	주관 개발
북한강일원 댐 유역 오염원조사 및 수질, 수생태 관리 방안 수립	수자원공사	'17.06. ~ '17.12.	데이터 분석	컨설팅
글로벌 재난안전 리스크 정보 탐색 및 모니터링 기술 개발	국립재난안전연구원	'17.06. ~ '17.12.	재난정보 분석 시스템 개발	시스템 개발
클라우드 기반의 인지컴퓨팅 시스템 구축을 위한 컨설팅 서비스	IBM	'16.08. ~ '16.11.	분석 컨설팅	주관 개발
삼성페이 서비스 고도화를 위한 분석 시스템	삼성전자	'16.03. ~ '16.06.	분석 컨설팅, 솔루션 납품	분석 컨설팅
재난안전 맞춤형 텍스트 탐색 및 실시간 분석 UI 개발	국립재난안전연구원	'16.06. ~ '16.12.	재난정보 분석 시스템 개발	시스템 개발
삼성생명 컨설팅 영업지원 시스템	삼성생명	'15. 6. ~ '15.12.	머신러닝 시스템 개발	컨소시엄 개발
삼성생명 BDA 경영자원화 구축	삼성생명	'15.06. ~ '15.12.	솔루션 납품, 모듈 커스터마이징	컨소시엄 개발
텍스트마이닝을 이용한 교통정보체계 감사토픽 분석	감사원	'15.02. ~ '15.04.	시스템 개발	Term License
미래위험 변화예측을 위한 사회환경탐색기술 개발	국립재난안전연구원	'14.05. ~ '14.08.	텍스트마이닝, 토픽분석	시스템 개발
리얼상품평 프로젝트 언어처리 모듈 구축	GS홈쇼핑	'12.02. ~ '12.08.	리뷰 자동 분석 시스템 개발	솔루션 도입