#### 하둡

HDFS 일자별 디렉터리 수집기

HDFS를 이용하여 일자별로 생성되는 디렉터리를 순회하면서 파일 이름 패턴과 일치하는 모든 텍스트 로그 파일을 수집합니다.



* HDFS 사이트: 미리 설정한 HDFS 사이트 이름을 지정합니다. 시스템 설정 메뉴의 HDFS 프로파일 메뉴에서 설정할 수 있습니다.
* 모니터링 기간: 실시간으로 파일 변화 모니터링할 기간을 설정합니다. 0으로 설정 시 실시간 수집을 비활성화 합니다.
* 디렉터리 경로: 로그 파일이 위치하는 HDFS 경로를 의미합니다.
* 파일이름 패턴: HDFS 디렉터리 경로에 존재하는 파일 중 파일 이름이 정규표현식 패턴에 일치하는 경우에만 수집합니다. 정규표현식 그룹을 쓰는 경우 파일 이름에서 날짜 문자열을 추출합니다.
* 디렉터리 날짜 정규표현식: 디렉터리에서 날짜 문자열을 추출합니다. 괄호(()) 문자를 통해 정규표현식 그룹으로 묶인 모든 부분을 이어붙여서 하나의 날짜 문자열을 만들어냅니다.
* 디렉터리 날짜 포맷: 날짜 문자열을 파싱하는데 사용할 날짜 포맷을 설정합니다. (예: yyyy-MM-dd HH:mm:ss)
* 디렉터리 날짜 사용 유무: 로그 생성 날짜를 디렉터리 이름에서 추출할지 여부를 설정하며 true 혹은 flase를 입력합니다. 미설정 시 false를 기본값으로 사용합니다.
* 과거 로그 수집 시작 일자: yyyyMMdd 포맷으로 시작 일자를 지정합니다. 미설정 시 과거 로그 수집이 비활성화됩니다.
* 과거 로그 수집 끝 일자: yyyyMMdd 포맷으로 끝 일자를 지정합니다. 미설정 시 과거 로그 수집이 비활성화됩니다.
* 날짜 정규표현식: 로그에서 날짜 문자열을 추출합니다. 정규표현식 그룹으로 묶인 모든 부분을 이어붙여서 하나의 날짜 문자열을 만들어냅니다.
* 날짜 포맷: 날짜 문자열을 파싱하는데 사용할 날짜 포맷을 설정합니다. (예: yyyy-MM-dd HH:mm:ss)
* 날짜 로케일: 날짜 문자열의 로케일. 가령 날짜 파싱 포맷의 지시자 중 MMM의 해석은 로케일에 따라 “Jan” 혹은 “1월”로 해석됩니다. 기본값은 en입니다.
* 시간대: 날짜 문자열의 시간대를 입력합니다. (예: KST 또는 Asia/Seoul)
* 로그 시작 정규표현식: 로그의 시작 부분을 인식하는 정규표현식을 지정합니다. 멀티라인 로그의 경우에 사용되며, 지정하지 않으면 줄 단위로 읽어들입니다.
* 로그 끝 정규표현식: 로그의 끝 부분을 인식하는 정규표현식을 지정합니다. 멀티라인 로그의 경우에 사용되며, 지정하지 않으면 줄 단위로 읽어들입니다.
* 문자 집합: 텍스트 파일 해석에 사용할 문자집합 코드를 입력합니다. 기본값은 utf-8입니다.
* 파일이름 태그: 수집한 파일명을 태깅할 필드 이름을 입력합니다.
* 파일 끝 수집 활성화: 로그 끝이 식별되지 않더라도 파일 끝에 도달하면 로그를 수집합니다. "true"로 입력할 경우 활성화되며, 미설정시 "false"를 기본 값으로 비활성화 됩니다.
* 개행 문자로 CR 사용: 캐리지 리턴 코드를 줄바꿈 문자로 사용합니다. "true"로 입력할 경우 활성화되며, 미설정시 "false"를 기본 값으로 비활성화 됩니다.
* 트랜스포머 설정: 수집된 로그에 적용할 트랜스포머를 선택합니다. "선택하세요."로 설정할 경우 트랜스포머가 적용되지 않고 원본 그대로 수집합니다. 트랜스포머에 대한 자세한 내용은 다음 항목을 참고하세요. 12.2.2 트랜스포머

HDFS 텍스트 파일 수집기

HDFS 사이트 설정이 완료된 후, 로그 수집 설정을 통해 HDFS에 적재되는 파일을 실시간으로 수집할 수 있습니다. 다른 프로세스에서 HDFS 파일을 아직 닫지 않았거나 마지막 HDFS 블럭이 디스크로 플러시되지 않은 상태이더라도, 해당 데이터를 읽어들일 수 있습니다.

HDFS 텍스트 파일 수집 (hdfs-text) 로거는 HDFS에서 읽어온다는 점을 제외하면 디렉터리 로그 파일 수집 설정과 동일합니다.



* HDFS 사이트: 미리 설정한 HDFS 사이트 이름을 지정합니다. 시스템 설정 메뉴의 HDFS 프로파일 메뉴에서 설정할 수 있습니다.
* 디렉터리 경로: 로그 파일이 위치하는 HDFS 경로를 의미합니다.
* 파일이름 패턴: HDFS 디렉터리 경로에 존재하는 파일 중 파일 이름이 정규표현식 패턴에 일치하는 경우에만 수집합니다. 정규표현식 그룹을 쓰는 경우 파일 이름에서 날짜 문자열을 추출합니다.
* 날짜 정규표현식: 로그에서 날짜 문자열을 추출합니다. 정규표현식 그룹으로 묶인 모든 부분을 이어붙여서 하나의 날짜 문자열을 만들어냅니다. 파일이름 정규표현식의 그룹으로 추출된 날짜문자열은 가장 앞 부분에 위치합니다.
* 날짜 포맷: 날짜 문자열을 파싱하는데 사용할 날짜 포맷을 설정합니다. (예: yyyy-MM-dd HH:mm:ss)
* 날짜 로케일: 날짜 문자열의 로케일. 가령 날짜 파싱 포맷의 지시자 중 MMM의 해석은 로케일에 따라 “Jan” 혹은 “1월”로 해석됩니다. 기본값은 en입니다.
* 시간대: 날짜 문자열의 시간대를 입력합니다. (예: KST 또는 Asia/Seoul)
* 로그 시작 정규표현식: 로그의 시작 부분을 인식하는 정규표현식을 지정합니다. 멀티라인 로그의 경우에 사용되며, 지정하지 않으면 줄 단위로 읽어들입니다.
* 로그 끝 정규표현식: 로그의 끝 부분을 인식하는 정규표현식을 지정합니다. 멀티라인 로그의 경우에 사용되며, 지정하지 않으면 줄 단위로 읽어들입니다.
* 문자 집합: 텍스트 파일 해석에 사용할 문자집합 코드를 입력합니다. 기본값은 utf-8입니다.
* 파일이름 태그: 수집한 파일명을 태깅할 필드 이름을 입력합니다.
* 트랜스포머 설정: 수집된 로그에 적용할 트랜스포머를 선택합니다. "선택하세요."로 설정할 경우 트랜스포머가 적용되지 않고 원본 그대로 수집합니다. 트랜스포머에 대한 자세한 내용은 다음 항목을 참고하세요. 12.2.2 트랜스포머